

AD _____

Award Number: DAMD17-98-1-8045

TITLE: Improving Clinical Diagnosis Through Change Detection in Mammography

PRINCIPAL INVESTIGATOR: Yue-Joseph Wang, Ph.D.

CONTRACTING ORGANIZATION: The Catholic University of America
Washington, DC 20064

REPORT DATE: March 2003

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20030829 038

REPORT DOCUMENTATION PAGE

*Form Approved
OMB No. 074-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED	
	March 2003	Annual Summary (1 Sep 98 - 28 Feb 03)	
4. TITLE AND SUBTITLE		5. FUNDING NUMBERS	
Improving Clinical Diagnosis Through Change Detection in Mammography		DAMD17-98-1-8045	
6. AUTHOR(S): Yue-Joseph Wang, Ph.D.			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The Catholic University of America Washington, DC 20064 E-MAIL: wang@pluto.ee.cua.edu		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012		10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited			12b. DISTRIBUTION CODE
13. ABSTRACT (Maximum 200 Words) Temporal change of mass lesions overtime is a key piece of information in computer-aided diagnosis of breast cancer and treatment monitoring. For a specific patient, change detection is a critical step to select lesion candidates for follow-up diagnosis performed by either clinicians or computers. The purpose of the project is to develop an automatic change detection method to quantitatively extract the clinically important changes of suspicious lesions, upgrade the existing CAD system, and thus improve the clinical diagnosis of breast cancer. In particular, we have developed (1) PAR/mPAR/MLP/TPS based hybrid registration software to align sequential mammograms involving non-rigid deformation; (2) site model based change detection scheme to detect new lesions and/or select lesion candidates for further computer analysis; (3) feature extraction algorithm to obtain discriminative imagery features of true masses against mass-like normal tissues; and (4) neural network based decision support system(s) for mass detection. Image registration algorithm effectively recovers non-rigid deformation between mammograms. Site model based change detection scheme automatically detects the subtle changes and prioritizes lesion candidates with significant changes. The combined change-triggered and appearance-triggered initial lesion candidate selection increases the sensitivity of existing CAD system. The neural network based classifiers further improve the specificity of mass detection. The performance was initially 0.78-0.80 for the areas A_z under the ROC curves using the conventional neural network, and later improved to A_z values of 0.84-0.89 when using the newly developed multiple circular path neural networks.			
14. SUBJECT TERMS: breast cancer, change detection, mass detection, computer-aided diagnosis			15. NUMBER OF PAGES 166
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited

3. TABLE OF CONTENTS

Front Cover	1
SF 298	2
Table of Contents	3
Introduction	4
Body	5
Key Research Accomplishments	19
Reportable Outcomes	19
Conclusions	21
References	21
Appendices	21

4. INTRODUCTION

Temporal change of mass lesions overtime is a key piece of information in computer-aided diagnosis of breast cancer and treatment monitoring, the **purpose** of the project is to develop an automatic change detection method to quantitatively extract the clinically important changes of suspicious lesions, upgrade the existing CAD system, and thus improve the clinical diagnosis of breast cancer. We will build a site model for each individual patient for monitoring the breast tissue changes and extend our current research on image registration and change detection to the early detection of breast cancer. **Specific aims** include: 1) registration and segmentation of deformable breast tissue structures across a series of mammograms; 2) construction of a site model of the mammogram for individual patients showing the locations of regions of interest and associated diagnostic information; 3) identification of clinically significant changes in both global and local mass areas within the breast; and 4) integration and evaluation of the developed techniques with existing CAD prototype. At conclusion of this project, we anticipate **achieving** the following: 1) establish a reliable technique of monitoring breast tissue changes associated with cancerous masses; 2) deliver a CAD prototype that can incorporate tissue change information from additional mammograms; 3) evaluate the merit of combining change detection and CAD for improved clinical diagnosis using multiple mammograms; and 4) acquire the experience necessary to explore multimodality imaging for unified detection, diagnosis and treatment assessment of breast cancer.

5. BODY-Final Summary

5.1 Statement of Work

This project aims to develop and integrate site-model based change detection with an improved CAD system for the purpose of clinical use. Our technical emphasis will be in the combined methods of using spatial and temporal data of the mammographic images and based on pattern recognition power of the site model based change detection and defined features as those indicated by ACR BI-RADS lexicons. Specific tasks and time line of this project are listed as follows:

- Months 1-45: Development of mammography database: (a) collection of cases, (b) establishment of patient record, (c) digitization of mammograms, and (d) computer archival for the digitized mammograms.
- Months 1-6: (a) Morphological studies for the background reduction and mass enhancement.
(b) Analysis of the mass features which can be used for differentiation from false masses.
- Months 7-12: (a) Image registration and site model construction for individual patients.
(b) Extraction of suspected mass areas using region growing and valley blocking techniques.
(c) Annual report.
- Months 13-18: (a) Continuous work on extraction of suspected mass areas using region growing and valley blocking techniques.
(b) Development of multilayer perceptron neural network (MLPNN).
(c) Development of convolution neural network (CNN) as computerized vision system for mass detection.
- Months 19-24: (a) Development of the dual target convolution neural network (DTCNN).
(b) Annual report.
- Months 25-30: (a) Initial laboratory test of the MLPNN.
(b) Initial laboratory test of the DTCNN.
(c) Development of the combined neural network system based on the results of the initial tests.
- Months 31-36: (a) Evaluation of initial results from the MLPNN and DTCNN.
(b) Laboratory test of the combined neural network system. The combined system can be trained by:
 (a.1) Combining the results from the MLPNN and DTCNN.
 (a.2) Performing further training with the trained MLPNN and DTCNN.
 (a.3) Performing further training from scratch.
(c) Annual report.
- Months 37-42: (a) Development of CAD system that can quantify the lesion changes over time.
(b) FROC study for laboratory test of the system.
- Months 43-48: (a) Performing a simulated clinical study with a small scale ROC study. Comparing the results with and without the assistance from developed CAD.
(b) Analysis of the studies.

(c) Final report.

5.2 Detailed Report

The detection and tracking of masses from mammograms taken from different views and over a period of time and determination of the changes in shape and size of lesions can provide vital clues for diagnosis and treatment assessment. This requires accurate fusion of mammograms taken over a period of time and emphasizing on the change of masses over the time. The unchanged masses represent a group with much lower potential of being cancer. The longer they have been unchanged, the greater the likelihood that they are benign. Three groups of suspicious masses will be tracked and highlighted for closer clinical inspection and for further CAD analysis. Based on a series of mammograms, the suspicious masses detected within the common overlapping area will be identified and the one-to-one correspondence of each pair will be established. The unchanged masses among them represent a group with possible less clinical potential in developing breast cancer. The masses observed in non-overlapping areas will be carefully analyzed to determine if they are the new lesions or the missed masses in some of these mammograms. Finally, the changed masses including new lesions will be quantitatively characterized to provide accurate input to the radiologists or the follow-on components of CAD procedure, since they are the clinically significant signs of breast cancer.

5.2.1 Data Acquisition

The acquired database consists of three data sets of breast cancer images. In the first data set, we collected 200 mammograms from the MIAS database and the BAMC database. Of the 200 mammograms, 50 mammograms are normal, and each of the remaining 150 mammograms contains at least one mass of varying size, subtlety, and location. Both the cranio-caudal (CC) and medio-lateral oblique (MLO) projection views were used. The films were digitized with a computer format of $2048 \times 2500 \times 12$ bits (for an $8'' \times 10''$ area where each pixel represents 100 um square). (See attached papers #1-4 for more detail description.)

The second data set consists of sequential mammograms of 6 patients taken over a period of time between 1996 and 1999. Both the CC and MLO projection views were used. The films were digitized with a computer format of $2048 \times 2500 \times 12$ bits. (See attached paper #6 for more detail description.)

The third data set contains three-dimensional (3D) sequential breast images obtained by dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) of both patients and rats. Each image plane is $512 \times 512 \times 12$ bits.

We have established in-house computer archival for all three data sets that also includes the corresponding related clinical record. The database is comprised of cases selected using the following criteria: (a) the database should include all types of breasts, such as fatty, dense, and moderate tissue breasts, (b) the database should include various sizes of masses, (c) the database should cover all types of masses such as round, oval, lobulated, irregular, etc., (d) at least 20% of masses in the database should be malignant, (e) the database should equally include cases both with and without masses, and (f) the cases composing the database should be selected without regard to race. At GUMC, the race population of patients with access to breast imaging is approximately 28% Black, 56% White, 10% Asian, and 6% Hispanic women. Each case consists of 4 old and 4 new mammograms. The determination of old and new mammograms is based on the date of examinations.

5.2.2. Image Background Correction and Mass Lesion Enhancement

One of the main difficulties in automatic mass-detection is that mammographic masses are often overlapped with breast tissues. In such cases, it is necessary to remove bright background caused by breast tissues but to keep mass-signals. For this purpose, background correction is an indispensable technique for mass detection.

The theory of mathematical morphology is powerful in analyzing and describing geometrical relations. Essentially it is a formalization of intuitive concepts such as size or shape. The two basic morphological operations are “erosion” and “dilation,” which are consistently defined for binary and gray-scale images. Using these two basic operations, two other basic and important operators, “opening” and “closing”, can be defined as follows:

$$\text{opening: } X_B \equiv (X \ominus B) \oplus B, \quad (1)$$

$$\text{closing: } X^B \equiv (X \oplus B) \ominus B, \quad (2)$$

where X indicates the original image, B represents the structuring element, and \oplus and \ominus indicate the operations “dilation” and “erosion,” respectively. Based on the “opening” operation, we have developed an operation for background correction. The operation is represented by

$$X - X_B = X - (X \ominus B) \oplus B. \quad (3)$$

This equation represents the subtraction of the image processed by the operator “opening” from the original image.

Figure 1 shows the effect of the operation represented by equation (3): (A) illustrates a structuring element, (B) shows the original signal (gray line) and the processed signal (black line) by “opening”, and (C) denotes the final output signal of the morphological operation. The final profile in (C) was obtained by subtracting the black profile signal from the gray profile signals in (B). Note that the detected peak signals were not affected by the operation. Hence the mass signals detected by the operation retain their original shapes.

As can be seen in this graph, the size of the detected peak significantly depends on the size of the structuring element. All peaks, which are smaller than the structuring element, can be detected. In our mass detection process, a 52 pixel-diameter structuring element will be used to detect masses whose sizes are less than 52 pixels in diameter. An object with a diameter of 52 pixels in a 512×625 pixel reduced image occupies 250 pixels in its original digitized image, and its real size is expected to be about 2.5 cm.

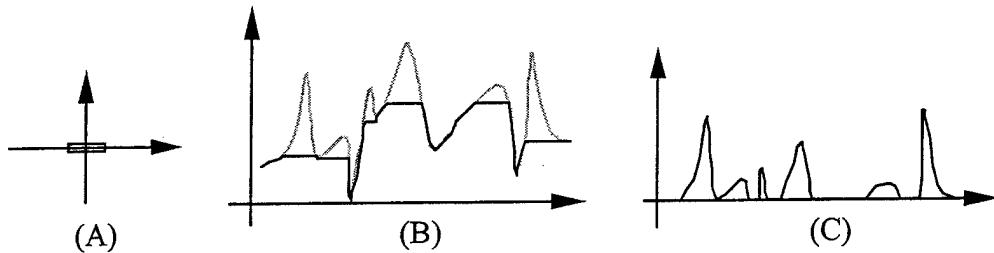


Figure 1. An example of the morphological operation: (A) structuring element, (B) original signal (gray line) and signal after opening (black line), and (C) output signal of the morphological filtering.

See attached paper #2, Section II and Figure 2, and paper # 4, Section III-A, for more detail descriptions.

5.2.3 Discriminatory Feature Extraction

Most commonly, breast cancer presents as a mass. The same lesion shows a somewhat different picture from one projection to the other. Difficulties in mass detection also vary with the underlying breast parenchyma. In the fatty breast, masses are generally easy to detect. With the dense breast, mass detection is more difficult and auxiliary signs aid this detection. Breasts can contain one, several, or many masses. When there is one mass, the decision process is based on its size, shape, and margins. The larger the mass is and the less well-defined its margins, the greater the chance of cancer. When there are several masses, one looks at each, trying to determine whether any has features to suggest cancer (poorly defined,

spiculate, unusually radiodense for size) and one also looks to see whether any mass is different in appearance from the others. Multiple small, well-defined, similar masses presenting bilaterally are all likely to be benign. The greater the asymmetry, size, lack of circularity, edge unsharpness, and radiodensity, the more suspicious. In this study, we used several computational features highly associated with four major features of breast masses routinely used in clinical reading:

Density - Malignant lesions tend to have greater radiographic density due to high attenuation and less compressibility of cancer than normal tissue. Radiolucent lesions are typically benign and the diagnosis can be made from the mammogram.

Size - If the lesion has morphological features suggesting malignancy, it should be considered suspicious regardless of the size. Isolated masses with non-cystic densities greater than 8 mm in diameter can be malignant. In general, the larger a lesion, the more suspicious it is.

Shape - The more irregular the shape of a lesion, the more likely the possibility of malignancy. Lesions tend to be round, ovoid and/or lobulated. Small and frequent lobulations are suspicious. Lesions in the lateral aspect of the breast near the edge of the parenchyma with a reniform shape and a hilar indentation or notch usually represent a benign intramammary lymph node. Breast carcinoma hidden in the dense tissues can cause parenchymal retraction, which possess different shapes.

Margins - The margins of the lesion should be carefully evaluated for areas of spiculation, stellate patterns or ill-defined regions. Most breast cancers have ill-defined margins secondary to tumor infiltration and associated fibrosis. The appearance of spiculations and a more diffuse stellate pattern are almost pathognomonic for cancer. Lesions with sharply defined margins have a high likelihood of being benign; however, up to 7% of malignant lesions can be well circumscribed.

These are known clinical features and have been adapted in "Breast Imaging - Reporting and Data System" (BI-RAD) of the American College of Radiology (ACR).

Feature extraction methods have played essential roles in many pattern recognition tasks. Once the features associated with an image pattern are extracted accurately, they can be used to distinguish one class of patterns from the others. Recently, many investigators have found that the multilayer perceptron neural network using the error back propagation training technique is a very powerful tool to serve as an analyzer (or classifier). Recently, the back propagation neural network (BPNN) for classification of features has widely been used in the field of computer-aided diagnosis.

The success of using an analyzer for a pattern recognition task would rely on two issues: (a) selected features that could describe discrepancy between patterns and (b) accuracy of the feature computation. Should either one fail, no analyzer or classifier would be able to achieve the expected performance. By analyzing many clinical samples of various sizes of masses, we found that the peripheral portion of the mass plays an important role for mammographers to make a diagnosis. The mammographer usually evaluates the surrounding background of a radiodense area when breast cancer is suspected.

We, therefore, performed boundary detection of the suspected masses on the morphologically enhanced mammogram. A region growing with valley blocking technique was employed to delineate all the suspected areas. Then, the boundary was divided into 36 sectors (i.e., 10^0 per sector) using 36 equi-angle dividers radiated from the center of suspicious area. The following features were computed within each 10^0 sector of the area:

- (a) "l" - the length from the center of mass to the shortest boundary segment.
- (b) "a" - the normal angle of the boundary segment (or the value of $\cos(a)$).
- (c) "g" - the average gradient of gray value on the segment along the radial direction.

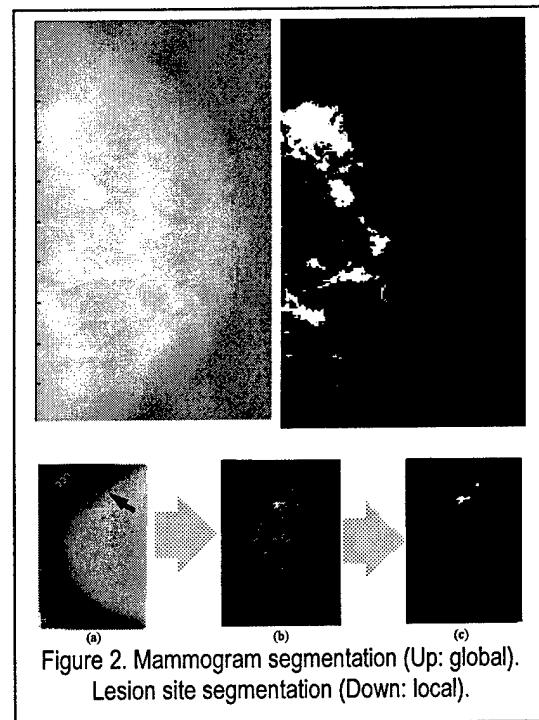


Figure 2. Mammogram segmentation (Up: global).
Lesion site segmentation (Down: local).

Technically speaking, this set of gradient values may also serve as a fuzzy system for the input layer in the neural network to be described.

(d) "c" - the gray value difference (i.e., contrast) along the radial direction. Averaged gray value (h_1) calculated from the mass area located at "1"/3 inside the boundary and the average background value (b_0) calculated from the peripheral area near "1"/3 outside of the suspicious area).

Hence, a total of 144 computed features (4 features/sector for 36 sectors) can be used as input values for the analysis of suspicious areas. The relationship between the computed features and BI-RADS descriptors are discussed below:

- (1) Mass Size - The 36 "1" values would provide sufficient data for the neural network to determine the size.
- (2) Mass Shape (round, oval, lobulated, or irregular) -The 36 "1" and 36 "a" values could approximate the shape of a mass.
- (3) Mass Margin (circumscribed, microlobulated, obscured, ill-defined, or spiculate) -

The 36 "g" and 36 "l" values should be able to describe the characteristics of the mass margin.

- (4) Mass Density (fat-containing, low density, isodense, or highly dense) -

The 36 "c" and 36 "g" values would be able to describe the density of the mass.

In short, the selected features are greatly associated with the main mass descriptors indicated in the BI-RADS. The reason for using 36 values for each nominated feature is four-fold: (a) mass boundary varies, it is difficult to describe an image pattern using a single value; (b) due to the general shape of the masses, the features of masses can be easily analyzed by the polar coordinate system; (c) in case some features are inaccurately computed in several directions due to the structure noises, such as the breast slender lines, there may still exist a sufficient number of correct features; (d) generally more accurate results can be produced by using subdivided parameters rather than using global parameters in a pattern recognition task. Other computational features (e.g., difference entropy and other higher order features) are eligible but require further investigation.

See attached papers #3 Section II and Table 1, #4 Section III-A, for detail descriptions.

5.2.4 Lesion Site Selection by Image Segmentation

We have analyzed mammograms that contain multiple anatomical objects through direct 2-D and/or 3-D tissue quantification and region segmentation. In particular, the tissue quantification is performed based on the standard finite normal mixture (SFNM) modeling of pixel image distribution, AIC and MDL guided model selection, and EM maximum likelihood model estimation. The region segmentation (e.g., the mass sites from the mammograms) is achieved based on inhomogeneous MRF modeling of context images and relaxation labeling of pixel memberships. Figure 2 shows a typical example. We have developed a new algorithm to perform lesion site segmentation as well as whole image segmentation. We have implemented the computer codes and pilot tested its effective applications to the digital phantoms, mammograms, and DEC-MRI images. The algorithm includes dual morphological filtering for signal enhancement and statistical model based tissue quantification and lesion segmentation. Our results have indicated that all the suspected lesion sites were successfully detected and the areas were accurately segmented.

See attached paper #2 for detail descriptions.

5.2.5. Site Model Based Hybrid Image Registration

Based image analysis results and all available clinical diagnostic information, we pilot constructed a patient specific site model. This model is a mathematical formulation of multimedia scene information,

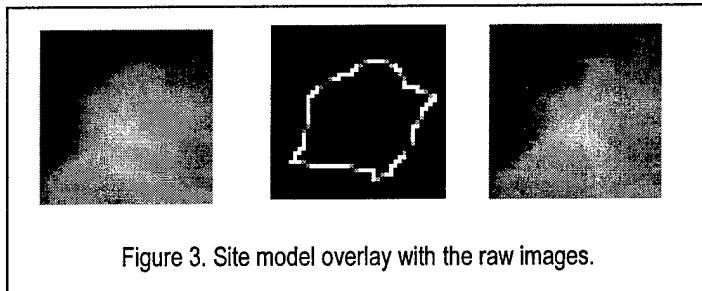


Figure 3. Site model overlay with the raw images.

mainly including object geometry (e.g., object location), reference labels (e.g., control points and/or objects), and expert's knowledge (e.g., lesion index, diagnostic comments, etc.).

Construction of a patient specific model (i.e., the site model) based on the outcome of image analysis includes objects, surface, and boundaries, of the normal tissues and detected/suspected lesions. This will provide a framework for (1) high accuracy change monitoring considering the patient variation and (2) effective data fusion incorporating prior/domain specific information.

Development of a multiple step algorithm for 2-D and 3-D image registration of image sequence data sets and multimodality image data sets. It consists of three major components: (1) principle axes registration (PAR), (2) site model support control feature alignment with localized PAR, namely mPAR, and (3) raw data matching via neural network based non-rigid warping between the two images, namely multilayer perceptron (MLP) and thin-plate spline (TPS). Figure 3 shows a result of local matching.

We have implemented a new hybrid registration algorithm aimed at the registration of non-rigid objects with minimal a prior knowledge, in which we have developed a methodology to combine multiple transforms together to determine a statistically composite geometric transform. The purposed algorithm combines rigid and non-rigid techniques to accomplish the registration tasks. The algorithm consists of two steps an initial step (rigid transform) which performs multi-object PAR registration where object correspondence is assumed known, and a final step (non-rigid transform) that uses thin-plate spline (TPS) based mapping where control point correspondence is determined via a detection and correspondence algorithm. The combination of these two steps is new and provides many advantages over existing methods. The first advantage is no requirement for point correspondence in the initial step. Only object correspondence is required which is usually much easier computationally to determine. True point correspondence is required at some point in the processing, but performing the determination after the image has been preliminarily aligned should allow for a more focused or narrow control point search windows because potential control points should now be closer spatially. The second advantage is the ability to model non-rigid transforms by considering each rigid transform as a piece wise component of a total non-rigid transform similar to modeling a non-linear function by linear pieces. This approach is a departure from traditionally registration approaches which usually follow either rigid or non-rigid transforms. In particular, we apply the combination method to multiple PAR transforms, but the method is generic and can be applied to any type of transform along as each cluster control point meets the particular requirement of the registration method in question. For example, to use an elastic registration method it is assumed we know the point correspondence of control points. In this algorithm, the image is assumed to contain several clustered control points, which follow a normal distribution, for which cluster correspondence is known (i.e. objects). The resulting transform now enables rigid transform methods to handle non-rigid transform assuming the clusters are sufficiently distributed through out image.

The registration process is supported by the concept of a site model and site model operations. The site model is a mathematical representation of a scene under analysis. A basic site model contains a geometric description of a scenes objects (area, size, and other attributes), raw data, and simple user input (previous tumor locations). The environment interacts with the site model through the site model operations: construction, image-to-site registration and model parameter update. The site model is constructed by thoroughly processing the first image in the sequence to obtain the parameters. The site

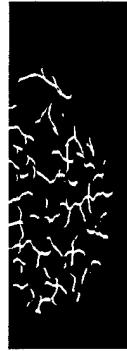


Figure 4. Control point extraction.

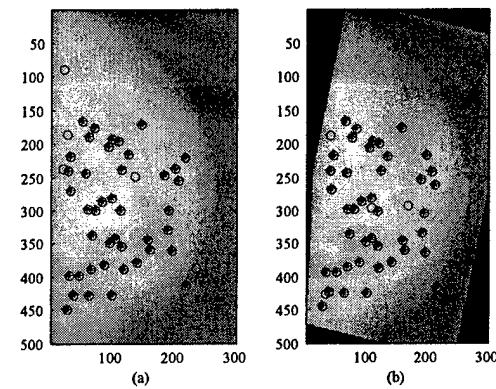


Figure 5. Control point matching.

model supports registration in three main ways. First, the site model forms the reference frame (reference image) for all subsequent images, thus allowing all of the images in the sequence to be alignment to a common coordinate system. Second, the model stores registration parameters like object contours, control points, and user identified regions. This effectively integrates both manual and automatic control objects in a single place. Third, the model stores previously detected change, this enables the current registration process to exclude the previously detected changed portion from the current analysis which improves algorithm robustness. In this research, we focus on the rigid, affine, and polynomial based registration methods to register the sequence of mammograms of the same patient. Image-to-site model registration is performed by a multi-step algorithm consisting of an initial and final phase. The initial phase registers the images using the principle axis of the skin line in conjunction with segmented internal objects to form a multi-object global rigid spatial-coordinate transform followed by a simple look up table for the intensity transform. The final registration phase consists of a global thin-plate spline transform derived from the control points of the interior breast tissue.

Figure 4 shows the result of control points extraction using our method. Figure 5 shows the corresponding control points in two similar breast phantoms. It can be seen that most control points are well matched using our PAR based initial registration. Figure 6 shows shows the corresponding control points in two real mammogram sequence. After our initial registration, stable control points are matched for further registration effort.

We have developed a neural computation based non-rigid registration methodology using multiple rigid transforms, in a piece-wise fashion, to model the registration process between images in a sequence. The registration methodology is a hybrid approach that combines registration without exact point correspondence via multi-object principal axes, and registration with point correspondence via polynomial transform. Neural computation is used, for the first, to combine the derived individual principal axes solutions for each object in a committee machine formulation, and to obtain the polynomial transform based on extracted control points using a multilayer perceptron (MLP).

In our method, we present a neural computation based non-rigid registration using piece-wise rigid transformation. The novel feature is to align two point sets without needing to establish explicit point correspondences, where the derivation is realized by minimizing the relative entropy between the two point distributions resulting in a maximum likelihood estimate of the transformation matrix. A committee

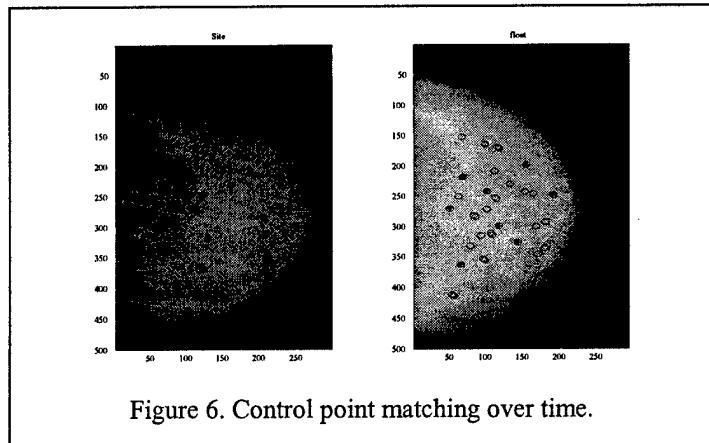


Figure 6. Control point matching over time.

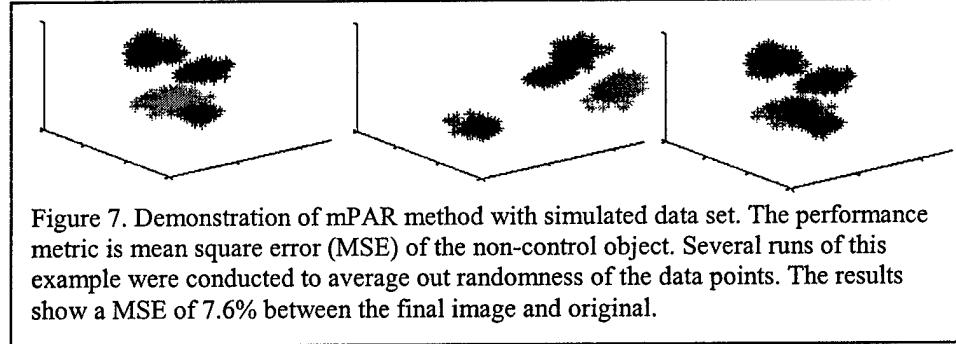


Figure 7. Demonstration of mPAR method with simulated data set. The performance metric is mean square error (MSE) of the non-control object. Several runs of this example were conducted to average out randomness of the data points. The results show a MSE of 7.6% between the final image and original.

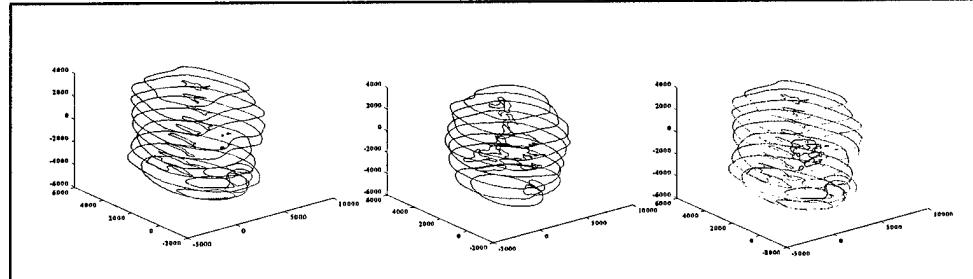


Figure 8. Demonstration of mPAR method with real prostate data set which contains a precise probabilistic map of prostate tumor distribution and the corresponding anatomic structure of a prostate.

machine approach is used for recovering the transformational geometry of the non-rigid structures. That is rather than using a single transformation matrix which gives rise to a large registration error, we attempt to interpolatively apply a mixture of transformations. By further generalizing PAR to a finite mixture registration (mPAR) scheme, with a soft partitioning of the data set, the mixture is fit using expectation-maximization (EM) algorithm. We then applied a probabilistic adaptive principal components extraction (PAPEX) algorithm, to estimate the transformational of the orthogonal set of eigenvalues and eigenvectors of the auto-covariance matrix. By applying a committee machine to a non-rigid registration, using FMR as the experts and PAPEX as a gating function, we can acquire the registration based on a mixture of piece-wise transformations of the data set. Then the correspondences control points are obtained. As a final step, the warped image is obtained using the neural network based non-linear mapping, to obtain the polynomial transform based on extracted control points using MLP.

Three examples are presented to demonstrate the techniques involved in the process, see Figures 7, 8, and 9. The first example uses four Gaussian clusters and focuses on the combination of the multiple transforms into a composite transform using finite mixture modeling techniques. The next examples present the complete process for prostate cancer registration and breast sequence analysis respectively. To verify performance, the results are compared to non-neural based implementations and other existing registration methods.

5.2.6. Site Model Based Change Detection

We have developed a patient specific site model concept to image-guided lesion monitoring. The site model was developed to monitor a site from a sequence of aerial images. In medical imaging, the site model idea was modified to accomplish application such as lesion monitoring, and disease detection. In addition, through update procedures the site model allows for the examination of the entire sequence together, to show region progression or to further highlight small changes. The main modification to the site model idea was the creation of another variable to store changes. In traditional site model formulations, new objects are added back into the image, but in the medical environment the site image is untouched. The changes are stored in the change map. The site image is untouched because it forms the base frame for comparison so any modification could alter results. Figure 10 shows a typical layout of the site model.

Change detection not only highlights existence of possible changed regions, but when combined with the site model provides a patient history by showing site progression. One of the key components of change detection is image registration. In this project, we applied our multi-step registration algorithm to mammogram sequences. Acceptable registration and change detection were obtained. Improvement in

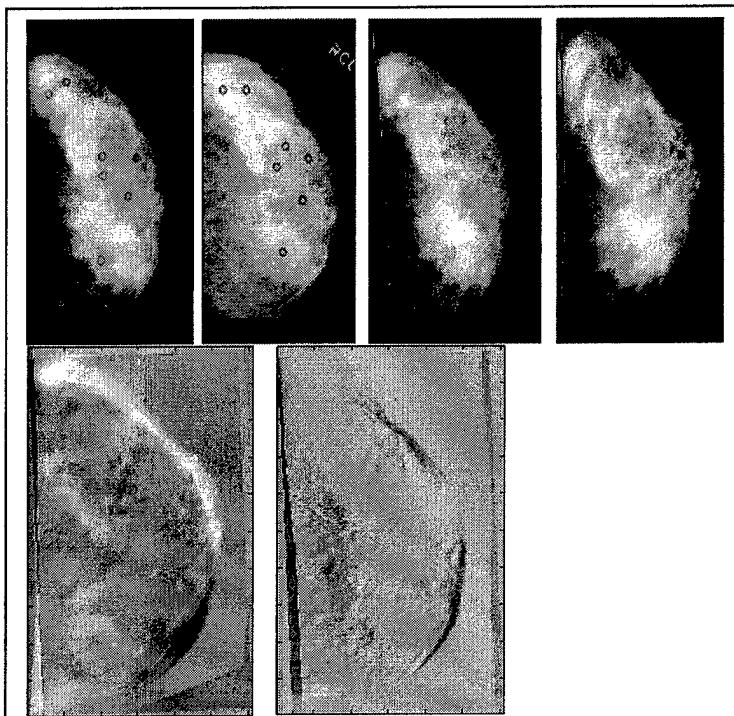


Figure 9. Difference images (lower row) after MLP and TPS image warping (up row).

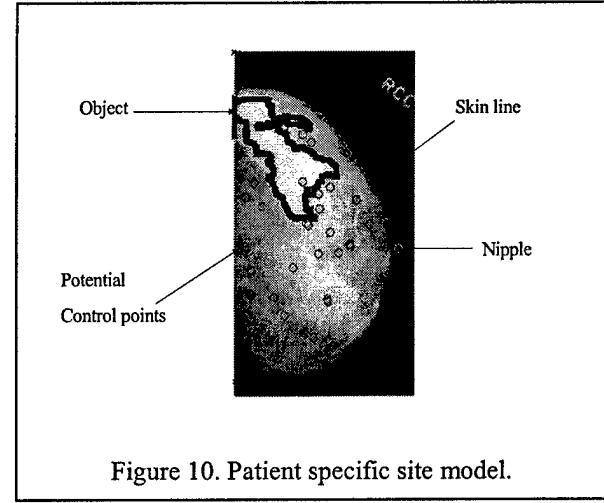


Figure 10. Patient specific site model.

control object selection and control point extraction would go along way to improving the overall results. The key to registration is landmarks between the images. In this research, we use objects and points as landmarks. Current methods of object and point selection are image dependent and ad hoc. Incorrect assignment of control points/objects could cause erroneous transformation. This change detection is not exact, but would be sufficient to flag a radiologist to review the area. The main results of this study consisted of the automatic alignment of mammograms, detection of change in a local window, and implementation of a mechanism to store and build up patient information via the site model.

Figures 11 and 12 show the results of automatic detection of local changes that could lead to the selection of new lesions.

This complete change detection algorithm was simulated with phantom images and real mammograms. The benefits of two steps in registration are apparent by looking at the mean square pixel error between no registration, single object PAR, and multi-PAR/TPS registration where the MSE drops almost 84% compared to only 70% with PAR alone. The change metric (joint global relative entropy (GRE)) was compared to two existing video sequence methods chi square and histogram difference. Joint GRE performed better as it was able to detect intensity changes, shift changes and shift/intensity changes. The quantification process estimated on average within 15% of the true objects size for the studies under considerations.

See attached papers #5 and #6 for more detail description.

5.2.7. Development of Neural Network Classifiers

In the clinical course of detecting masses, mammographers usually evaluate the surrounding background of a radiodense when breast cancer is suspected. In this study, we adapted this fundamental concept and computed features of the suspicious region in radial sections. These features were then arranged by circular convolution processes within a neural network, which led to an improvement in detecting mammographic masses.

In this study, randomly selected mammograms were processed by morphological enhancement techniques. Radiodense areas were isolated and delineated using a region growing algorithm with a valley blocking technique. The boundary of each region of interest was then divided into 36 sectors using 36 equi-angle dividers radiated from the center of the area. Four features at each section were computed: (1) the radius, (2) the normal angle of the boundary, (3) the average gradient along the radial direction, and (4) the gray value difference (i.e., contrast) along the radial direction. Hence, 144 computed features (i.e., 4 features per sector for 36 sectors) were used as input values for the newly invented multiple circular path neural network (MCPNN). The neural network is constructed to emphasize on the correlation information associated with the feature interactions within the angle and between adjacent angles.

We have tested this approach on our research database consisting of 91 mammograms. The overall performance in the detection of masses was 0.78-0.80 for the areas (A_z) under the ROC curves using the conventional neural network. Later, the performance was improved to A_z values of 0.84-0.89 using the multiple circular path neural networks.

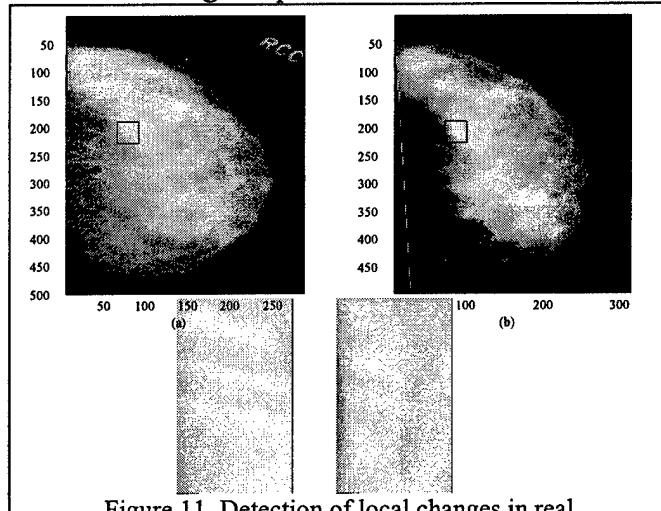


Figure 11. Detection of local changes in real mammograms over time.

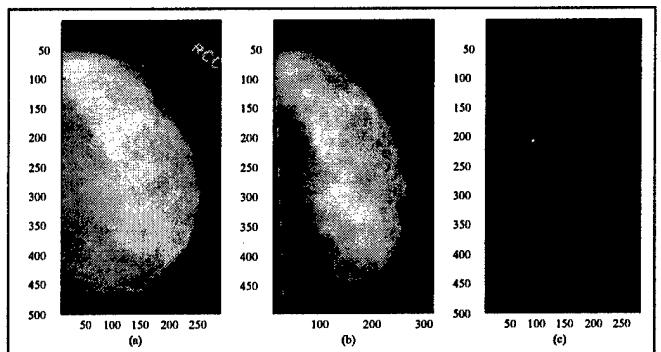


Figure 12 Change detection from real mammograms.

(A) Multiple paths with circular networking to instruct the neural network in analyzing sector features

We designed several neural network connections between the input and the first hidden layers as shown in Figure 13. Figure 13 (A), (B), and (C) illustrate the full connection, a self correlation (SC) networking, and a neighborhood correlation (NC) networking, respectively. Note that the input and hidden nodes should be completely matched when combining more than one path in the study. In this case, the correlation layers only function as branch connections between input and hidden layers. When using NC paths, networking engagement within multiple sectors (e.g., 20° , 30° , 40° , and 50° of the neighborhood correlation) can be grouped. The method of using the multiple correlation connections was motivated by our two-dimensional convolution neural network (2-D CNN) research experience where we found that more than 10 multiple convolution kernels were necessary to archive an outstanding neural network performance in the detection of lung nodules and microcalcifications.

Compared to 2-D CNN systems, the required computation using 1-D input features (i.e., 144) is relatively small. The combination of the networking paths described earlier for MCPNN was implemented using C programming language. The internal computation algorithm used in the MCPNN shares the same convolution process as that in the 2-D CNN. One additional training method using flipping invariance for 1-D convolution kernels was employed and is described in the section 3.3.(B).

The fully connected neural network is a standard back propagation neural network. The signals of the fully connected neural network join the other two network processes (SC and NC paths) at the single node of the output layer. The signal received at the output node is scaled between 0 and 1. During the training, 0 and 1 were assigned at the output node to perform back propagation computation for a non-mass and a mass, respectively. The back propagation was computed in such a way that the computed incremental errors (see equations (9) and (10) were retraced into three independent network paths (full-connected, SC, and NC paths). Besides the output layer, the SC and NC signals were independently arranged and are processed through two one-dimensional convolution processes in the forward propagation. The learning algorithms for all three paths were based on the backpropagation training method.

Let $N^0(n, s)$ represents input signal at the node n and sector s . The signal received at each node on the first hidden layer of the SC path is

$$N_{sc}^1(s) = \left(\sum_n N^0(n, s) \times W_{sc}(n) \right) + b_{sc}(s), \quad (4)$$

where $b_{sc}(s)$ represents the bias in s th sector. The signal gets into each node on the first hidden layer of the NC path is

$$N_{nc}^1(s) = \left(\sum_{s=-s1}^{s1} \sum_n N^0(n, s) \times W_{sc}(n, s) \right) + b_{nc}(s), \quad (5)$$

where $b(s)$ represents the bias in s th sector and $s1$ is 2 to cover -20 degree to 20 degrees of the fan. The signals in other hidden layers in each path are processed the same as the standard fully connected neural network. The output signal was collected from the last hidden layer and is given by,

$$O = S(N_p^l(n) \cdot W_p^l(n)), \quad (6)$$

where l denotes the hidden layer, p denotes the path, and $S(z)$ is a sigmoid function given by

$$S(z) = \frac{2}{1 + \exp(-z)} \quad (7)$$

The sigmoid function would produce modulated values ranging from 0 to 1.

Let the t -th change of the weight be $\Delta W_p^l(n)$ and the t -th change of the bias be $\Delta b_p^l(t)$. The error function is defined as

$$E = \frac{1}{2} (T - O)^2 \quad (8)$$

where T and O denote the desired output value and the actual output value, respectively when the input nodes $N^0(n, s)$, are entered in the network. In this model, the error back propagation algorithm, which updates the kernel weights, was given below:

$$\begin{aligned}\Delta W_p^l[t+1] &= \eta \left(\sum \delta_p^{l+1}(n, s) \cdot O_p^l(n, s) \right) + \alpha \Delta W_p^l[t] & (9) \\ \Delta b_p^l[t+1] &= \eta \sum \sum \delta_p^l(n, s) + \alpha \Delta b_p^l[t] & (10) \\ \delta_p^l(n, s) &= N_p^L(n, s) (\delta_p^{l+1}(n, s) \cdot W_p^l(n, s)) & (11)\end{aligned}$$

where $s = 0$ when $l \neq 0$. In the case of the last layer,

$$\delta_p^L = S'(N_p^L(n)) (T - O) \quad (12)$$

where $S'(z)$, η , α , and T denote the derivative of $S(z)$, the learning rate, the weighting factor contributed by the momentum term, and the desired output image, respectively.

During the training, we added an isotropic constraint to the weights in the 1-D convolution kernels and so that

$$W_p^0(n, -s) = W_p^0(n, s) \quad (13)$$

where p is not the fully connected path. These additional constraints were used to induce the kernels functioning as correlation processing filters and could facilitate the algorithm in searching for an appropriate linear filter.

(B) Training methods and the utilization of characteristics of flipping invariance of the features

Because we used the circular paths, there were no starting and ending sectors. The forward and back propagation computation can be started from any sector. Since the mass characteristics of the flipped patch remained the same, we flipped each patch in the training set and kept the same numerical value for the target output.

Since we designed a 10^0 increment for each rotation, each SC or NC networking would need to process through 36 times for the computed feature set for each image patch. To simplify this network computation, we shifted one small set (4 nodes) on the input layer a time to conduct the circular convolution process with the SC and NC kernels. By reversing the sequence of the sector, we can train the flipped version of the suspicious masses. Hence, the characteristics of flipping invariance literally increase the number of the training set by a factor of 2.

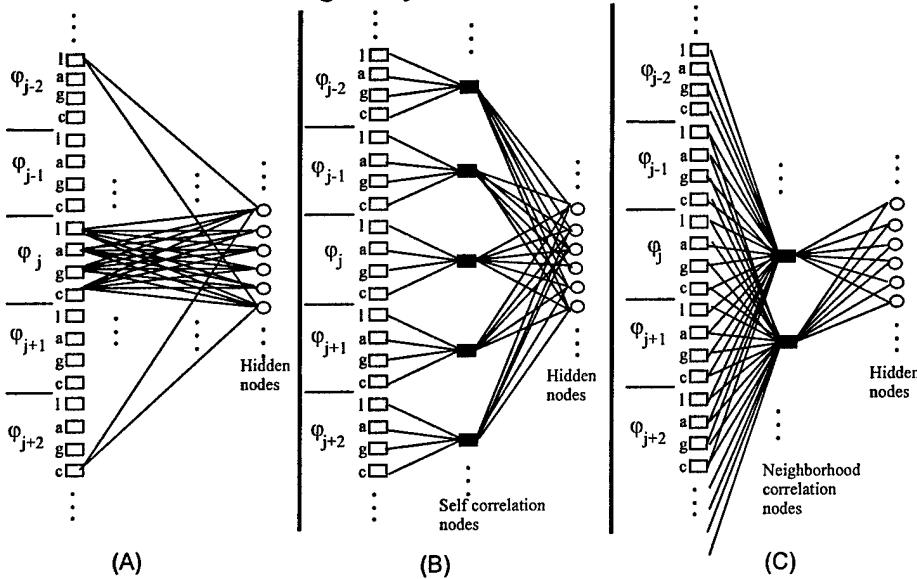


Figure 3. Three types of network paths connecting the input and the hidden layers:

- (A) Full connection.
- (B) A self correlation (SC) path; each node on the layer connects to a single set of the features (l,a,g,c) for the fan-in and fully connects to the hidden nodes for fan-out.
- (C) A neighborhood correlation (NC) path; each node on the layer connects to five adjacent sets of the features for the fan-in and fully connects to the hidden nodes for fan-out.

Note that the fan-in nets emphasizing self correlation in (B) and neighborhood correlation in (C) represent convolution weights (i.e., the same type of sectors possess the same set of weighting

factors).

We have described our approach on the feature extraction, the design of MCPNN, and its corresponding training method. Figure 14 shows a flow diagram of the proposed method. Since the MCPNN only alters the input data connection from the input to the first hidden layer, any learning algorithm can be applied within the neural network. For simplicity, we used the back propagation algorithm for both the conventional and proposed neural network systems in the following experiments.

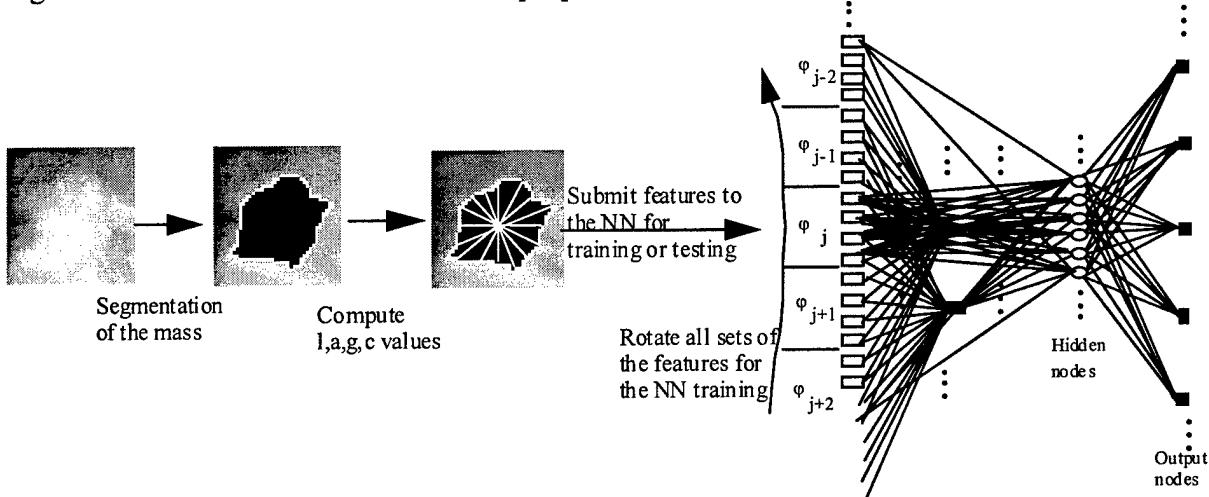


Figure 14. A schematic diagram, showing the MCPNN and sector features of masses, was used in the following study. See attached papers #1, #3, and #4, for more extended descriptions.

5.2.8. Pre-Clinical Evaluation of Upgraded CAD System and New Work on DCE-MRI

We selected 91 mammograms and digitized each mammogram with a computer format of $2048 \times 2500 \times 12$ bits (for an $8'' \times 11''$ area where each image pixel represents $100 \mu\text{m}$ square). No two mammograms were selected from the same patient film jacket. All the digitized mammograms were miniaturized to $512 \times 625 \times 12$ bits using 4×4 pixel averaging and were processed by the above methods to perform mass detection. Based on the corresponding biopsy reports, one experienced radiologist read all 91 mammograms and identified 75 areas containing masses. (Note that the reports recorded the malignancy of the biopsy specimens. The radiologist only used them as reference for the identification of masses.) Through the pre-process and the first step screen based on the circularity test, a total of 125 suspicious areas were extracted from the 91 digitized mammograms.

Experiment 1

We randomly selected 54 computer-segmented areas where 30 patches were matched with the radiologist's identification and 24 were not. This database was used to train two neural network systems: (1) a conventional 3-layer BP neural network (with 125 nodes in the hidden layer) and (2) the proposed MCPNN training method using the same neural network learning algorithm. The structure of the MCPNN was described earlier. However, we used one fully connected path, four SC paths, four 20° NC paths, four 30° NC paths, three 40° NC paths, and two 50° NC paths in the first step network connection for the MCPNN. All paths in the neural network have their hidden layers. Only one hidden layer per path was used. Both neural network systems were trained by the error back propagation algorithm by feeding the features from the input layer and registering the corresponding target value at the output side. Once the training of the neural networks was complete, we then used the remaining 71 computer segmented areas for the testing. None of the images and their corresponding patients in the testing set could be found in the training set. The neural network output values were fed into the LABROC program for the performance evaluation. The results indicated that the areas (A_z) under the receiving operator characteristic (ROC) curves were 0.781 and 0.844 using the conventional BPNN and the MCPNN, respectively. The ROC curves of these two neural network training methods are shown in Figure 15 (A).

We also invited another senior mammographer to conduct an ROC observer study. The mammographer was asked to rate each patch using a numerical scale ranging 0-10 for its likelihood of being a mass. These 71 numbers were also fed into the LABROC program. The mammographer's performance in Az on this set of test cases was 0.909. The corresponding ROC curve is also shown in Figure 15(A).

Experiment 2

We also conducted a leave-one-case-out experiment using the same database. In this experiment, we used those patches extracted from 90 mammograms for the training and used the patches (most of them are single) extracted from the remaining one mammogram as test objects. The procedure was repeated 91 times to allow every suspicious patch from each mammogram to be tested in the experiment. For each individual suspicious area, the computed features were identical to those used in Experiment 1. Again, both neural network systems were independently evaluated with the same procedure. The results indicated that the Az values were 0.799 and 0.887 using the conventional back propagation neural network and the MCPNN, respectively. Figure 15(B) shows the ROC curves of these two neural network systems using the leave-one-of-out procedure in the experiment.

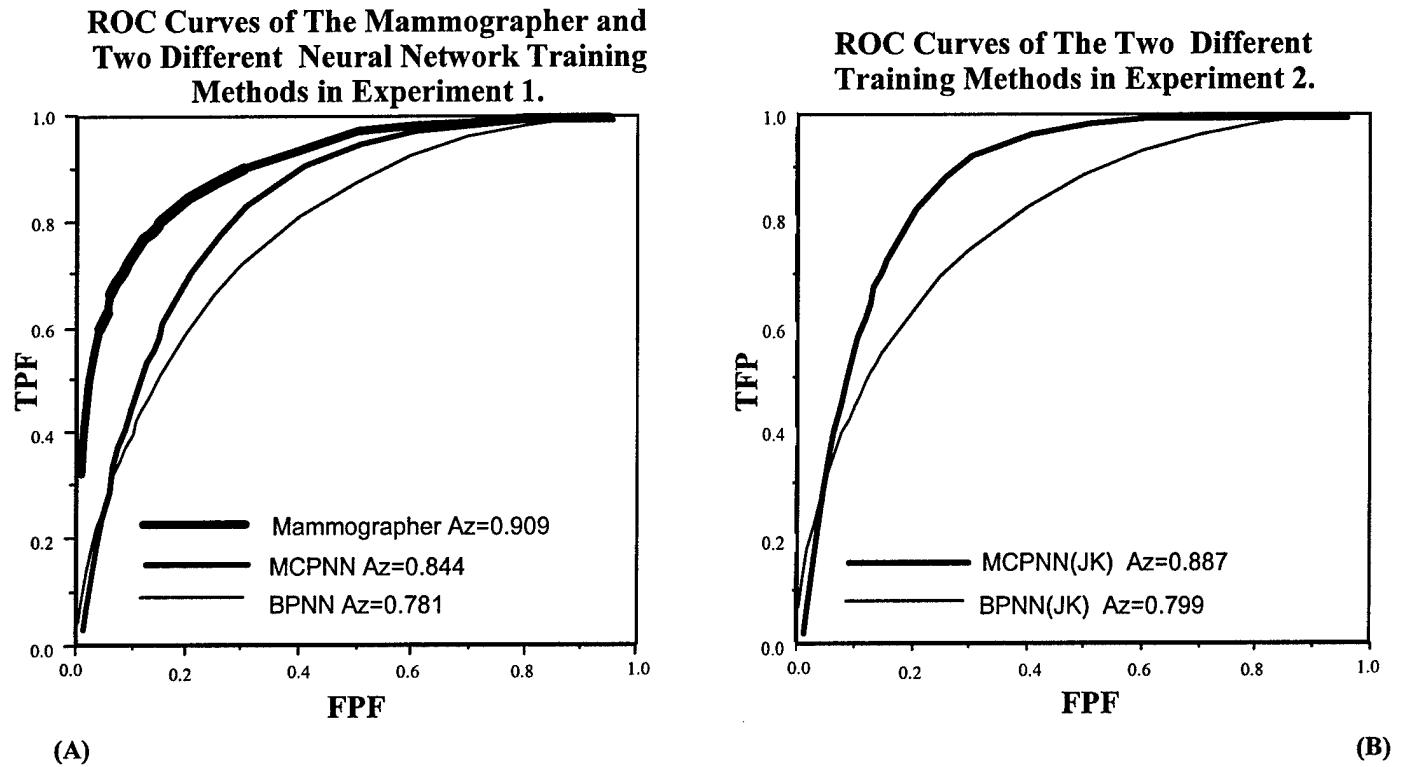


Figure 5. The ROC curves obtained from corresponding experiments.

- (A) The left figure shows that the performance of MCPNN training method is superior to that of the conventional input method. The highest curve is the ROC performance of the senior mammographer.
- (B) The right figure shows the ROC results with higher performance using the leave-one-case-out procedure as described in Experiment 2.

Through this study, we found that the selected features are somewhat effective in the detection of masses. These features were “computationally translated” from the qualitative descriptors of BI-RAD. Another uniqueness of this study was on the test of our newly developed MCPNN training method. In Experiment 1, we found that the performances of both neural network systems were increased. This might be due to the increased number of cases (from 54 to 124) in the training set. In Experiment 2, the Az value was improved by 0.043 using the MCPNN training method that was higher than Az difference

of 0.018 obtained by the conventional training method. The results implied that the MCPNN learned more effectively than the conventional BP when the number of training cases was increased.

It is known in the field of artificial intelligence that the key factors in pattern recognition are: (1) effective methods in the extraction of features and (2) analytic methods (e.g., back propagation neural network) for the extracted features. In this study, we showed that the training method designed to guide the analyzer is also an important factor to a success of a pattern recognition task. Though this finding is not new, the research of developing training methods for various pattern recognition tasks has not established in the field of medical imaging. In this work, we demonstrated that organized features with proper network connection and task-oriented guidance would assist the neural network in performing the task.

As far as the research in recognition of masses is concerned, we believe that main concept of using sectors is an effective approach. Note that any features arranged in the polar coordinate system can be trained by the MCPNN method. Since the MCPNN only coordinates the input data, the internal neural network learning algorithm can be changed to other learning algorithms. We believe that integration of effective feature and texture values computed at small sectors will be the research trend in mass detection. Current work focuses on neural network design and arrangement of features for effective pattern recognition of masses in medical imaging.

5.2.9. Extended Work on Static/Dynamic Contrast-Enhanced MRI

We have also extended our work to image registration of 3D sequential static/dynamic contrast-enhanced MRI breast images. The purpose is beyond mass detection and aims at quantitative assessment of breast cancer triggered local vascularization as a result of angiogenesis, as well as assessment of the responses to chemoprevention.

For 3D MRI breast image registration, PAR method has been initially used to register post-contrast image to pre-contrast image. The extracted skinline is used as control objects. The difference image shows the false regions of enhanced area of glandular tissue due to the misalignment. Figures 16 and 17 show the examples of such work. Based on initial registration by PAR and mPAR, we once again used a neural network MLP to refine the deformable warping. Figure 18 shows such work.

We have further developed blind source separation (BSS) method to characterize flow patterns associated with microvessel densities of breast cancers. Our method is based on newly developed partially-independent component analysis (PICA). Figure 19 illustrates the framework. Figure 20 shows the well-separated fast and slow flow patterns from dynamic contrast-enhanced MRI.

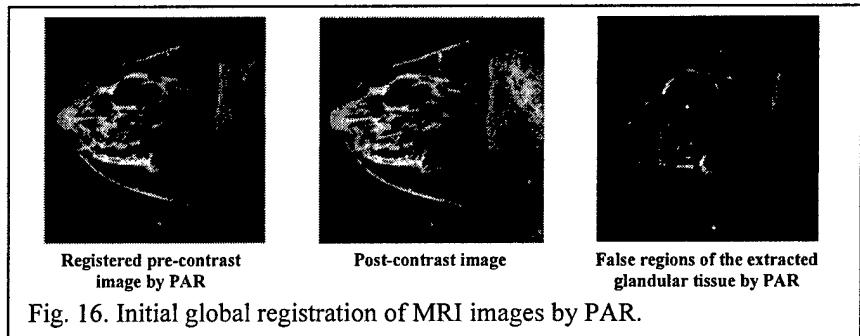


Fig. 16. Initial global registration of MRI images by PAR.

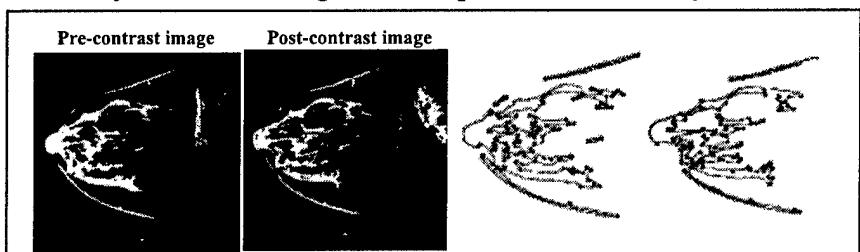


Fig. 17. Refined registration of MRI images by mPAR (left) and matching of extracted control point pairs for further warping (right).

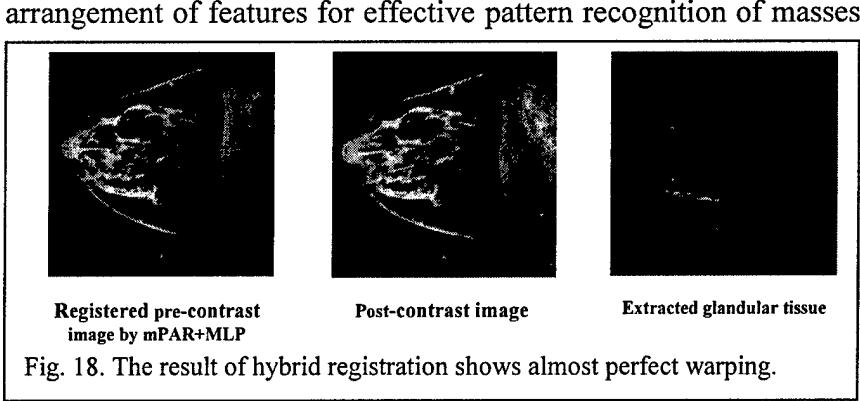


Fig. 18. The result of hybrid registration shows almost perfect warping.

5.2.10. Future Work

An integrated change detection and mass detection is a challenging task. There are three ways to incorporate change detection into CAD system. First, by comparing sequential images of same patient in screening, change detection can help detecting previously unrecognized/missed lesions. Second, in follow-up lesions, change detection can help to reduce false positive rate. Third, when change-related features can be accurately extracted, a modular classifier can be designed to combine change-derived diagnosis and shape/texture-derived diagnosis thus improve the specificity of mass detection. Such preliminary effort will also lead to multimodality CAD systems.

Another important area for future effort is image-based assessment of the responses to therapies. Our effort on using static contrast-enhanced MRI to assess the efficacy of chemoprevention to breast cancer has been promising.

6. KEY RESEARCH ACCOMPLISHMENTS

- We have proposed and developed effective methods to identify lesion sites automatically.
- We have proposed and developed an accurate and effective hybrid non-rigid image registration method to recover the deformations between images taken over a period of time.
- We have proposed and developed a patient-specific site model based method for quantitative change detection.
- We have proposed and developed a systematic method to extract discriminatory imagery features for mass detection using CAD methodology.
- We have proposed and developed various neural network based classifiers for mass detection after initial lesion candidates are identified and related features are extracted.
- We have tested our detection system using receiver operating characteristics (ROC) analysis. Our preliminary experiment has shown that new system outperform existing popular methods.

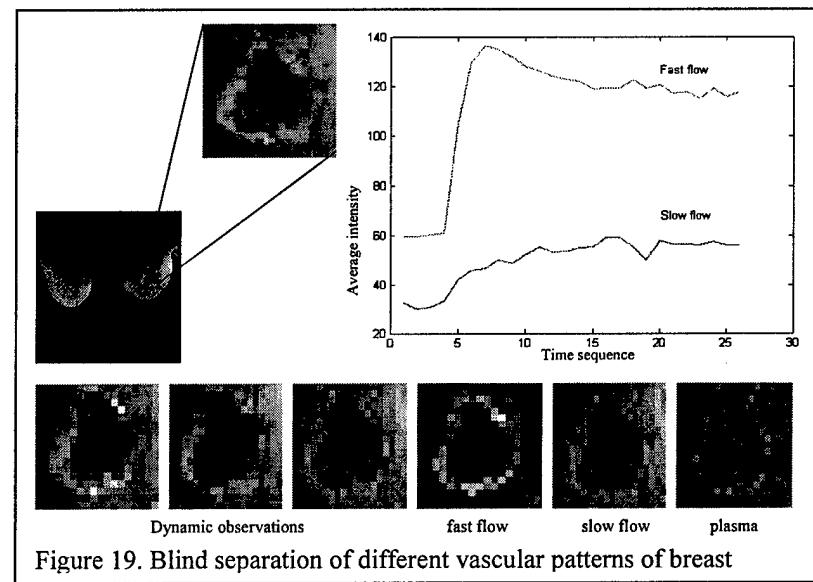


Figure 19. Blind separation of different vascular patterns of breast

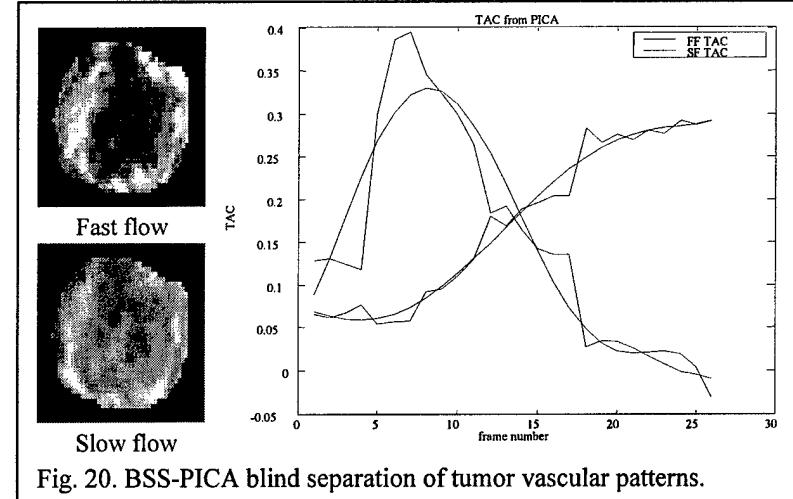


Fig. 20. BSS-PICA blind separation of tumor vascular patterns.

7. REPORTABLE OUTCOMES

Journal Publications

1. Y. Wang, S-H Lin, H. Li, and S-Y Kung, "Data Mapping by Probabilistic Modular Networks and Information Theoretic Criteria," *IEEE Transactions on Signal Processing*, vol. 46, no.12, pp. 3378-3397, December 1998.

2. H. Li, Y. Wang, K-J R. Liu, S-H B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection-Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 289-301, April 2001.
3. H. Li, Y. Wang, K-J R. Liu, S-H B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection-Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 302-313, April 2001.
4. S-C B. Lo, H. Li, Y. Wang, and M. T. Freedman, "A Multiple Circular Path Neural Network Architecture for Detection of Mammographic Masses," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 150-158, February 2002.
5. Y. Wang, K. Woods, and M. McClain, "Information-Theoretic Matching of Two Point Sets," *IEEE Transactions on Image Processing*, vol. 11, no. 8, pp. 868-872, August 2002.

Degree Granted

- Kelvin Woods, (2000) Ph.D. *Dissertation: Image Guided Diagnosis through Change Detection in Image Sequences*, The Catholic University of America. (**PI served as the major advisor**)
- Maxine McClain, (2000) M.S. *Thesis: Information Processing in Image-Guided Diagnosis and Therapy*, The Catholic University of America. (**PI served as the major advisor**)

External Funds Awarded

- Dues Technologies, Inc., Computer-aided Diagnosis for Lung Cancer Detection, 1999-present.
- Food and Drug Administration, Multidimensional Receiver Operating Characteristics (ROC) Analysis for Computer-Aided Diagnosis, 2001.
- Siemens Corporate Research, Inc., Computer-aided Diagnosis for Breast Cancer Detection, 2000.

Intergovernmental Personnel Agreement (IPA)

- Visiting Investigator, NIH Clinical Center, Radiology and Imaging Science Program, 2003-2004.

Proceedings Papers

6. K. Woods, L. Fan, C-W. Chen, and Y. Wang, "Model supported image registration and warping for change detection in computer-aided diagnosis," *Proc. 29th Applied Imagery Pattern Recognition Workshop*, pp. 180-186, Washington, DC October 2000.
7. Y. Wang, Z. Wang, L. Luo, S-H B. Lo, and M. T. Freedman, "Computer-based decision support system: visual mapping of featured database in computer-aided diagnosis," *Proc. SPIE Medical Imaging*, pp. 136-137, San Diego 2000.
8. K. Woods, Y. Wang, R. Srikanthana, and M. T. Freedman, "Model supported image registration for change detection in computer-aided diagnosis," *Proc. SPIE Medical Imaging*, pp. 1095-1106, San Diego 2000.
9. R. Srikanthana, K. Woods, J. Xuan, C. Nguyen, Y. Wang, "Non-rigid image registration by neural computations," *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 413-422, Falmouth, MA, September 2001.
10. R. Srikanthana, J. Xuan, M. Freedman, and Y. Wang, "Mixture of principal axes registration: a neural computation approach," *Proc. SPIE Medical Imaging*, vol. 4684, San Diego, CA, February 2002.
11. Y. Wang, J. Zhang, K. Huang, J. Khan, and Z. Szabo, "Independent component imaging of disease signatures," *Proc. IEEE Intl. Symp. Biomed. Imaging*, July 7-10, Washington, DC 2002.

List of Participants

Yue (Joseph) Wang, PhD, PI
Matthew T. Freedman, MD, Consultant
Ben Lo, PhD, Consultant
Kelvin Woods, Doctoral Student
Maxine McClain, Master Student
James Lu, Visiting Scholar
Teresa Osicka, Doctoral Student
Andy Srikanchana, Doctoral Student
Bin Yang, MS, Consultant
Zhiping Gu, PhD, Consultant

8. CONCLUSIONS

Through this Career Development Award (CDA), the PI has achieved his training in two aspects. Technically, we have: 1) established a reliable technique of monitoring breast tissue changes associated with cancerous masses; 2) delivered a CAD prototype that can incorporate tissue change information from additional mammograms; 3) evaluated the merit of combining change detection and CAD for improved clinical diagnosis using multiple mammograms; and 4) acquired the experience necessary to explore multimodality imaging (including function/molecular imaging) for unified detection, diagnosis and treatment assessment of breast cancer.

Professionally, the PI has gained much more knowledge and experience in breast cancer research. He has become a dedicated researcher on breast cancer research and has made differences in related areas. Based on this CDA training, the PI has broaden his research into microarray data analysis (e.g., bioinformatics, etc.) and function/molecular imaging for breast cancer research. See Attached Biosketch.

9. REFERENCES

N/A

10. APPENDICES

- 1) **Y. Wang**, S-H Lin, H. Li, and S-Y Kung, "Data Mapping by Probabilistic Modular Networks and Information Theoretic Criteria," *IEEE Transactions on Signal Processing*, vol. 46, no.12, pp. 3378-3397, December 1998.
- 2) H. Li, **Y. Wang**, K-J R. Liu, S-H B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection-Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 289-301, April 2001.
- 3) H. Li, **Y. Wang**, K-J R. Liu, S-H B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection-Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 302-313, April 2001.
- 4) S-C B. Lo, H. Li, **Y. Wang**, and M. T. Freedman, "A Multiple Circular Path Neural Network Architecture for Detection of Mammographic Masses," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 150-158, February 2002.
- 5) **Y. Wang**, K. Woods, and M. McClain, "Information-Theoretic Matching of Two Point Sets," *IEEE Transactions on Image Processing*, vol. 11, no. 8, pp. 868-872, August 2002.
- 6) **Y. Wang** and Kelvin Woods, Technical Report, TR-CUA001, 2000.
- 7) PI's Biosketch.

Data Mapping by Probabilistic Modular Networks and Information-Theoretic Criteria

Yue Wang, Shang-Hung Lin, Huai Li, and Sun-Yuan Kung, *Fellow, IEEE*

Abstract—The quantitative mapping of a database that represents a finite set of *classified* and/or *unclassified* data points may be decomposed into three distinctive learning tasks:

- 1) detection of the structure of each class model with locally mixture clusters;
- 2) estimation of the data distributions for each induced cluster inside each class;
- 3) classification of the data into classes that realizes the data memberships.

The mapping function accomplished by the probabilistic modular networks may then be constructed as the optimal estimator with respect to information theory, and each of the three tasks can be interpreted as an independent objective in real-world applications. We adapt a model fitting scheme that determines both the number and kernel of local clusters using information-theoretic criteria. The class distribution functions are then obtained by learning generalized Gaussian mixtures, where a *soft* classification of the data is performed by an efficient incremental algorithm. Further classification of the data is treated as a *hard* Bayesian detection problem, in particular, the decision boundaries between the classes are fine tuned by a reinforce or antireinforce supervised learning scheme. Examples of the application of this framework to medical image quantification, automated face recognition, and featured database analysis, are presented as well.

I. INTRODUCTION

THIS PAPER addresses the problem of mapping a database, given a finite set of data points (examples). The mapping function can therefore be interpreted as a quantitative representation of the contents (knowledge) contained in the database [1], [3], [4]. The data set may be a *classified* set, as in general clustering problems [2], [22], [25], it may be unclassified, as in unsupervised distribution learning [1], [12], [18], or it may be a partially classified set, as in pattern classification applications [5]–[7]. Instead of mapping the whole data set using a single complex network, in many applications, it is more practical to design a set of simple class subnets with locally mixture clusters, each one of which

Manuscript received July 24, 1997; revised January 5, 1998. This work was supported in part by the U.S. Army Medical Research and Materiel Command under Grant DAMD17-98-1-8046 and the National Institutes of Health under Grant 1R21RR12784-01. The associate editor coordinating the review of this paper and approving it for publication was Dr. Y. H. Hu.

Y. Wang is with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA (e-mail: wang@pluto.ee.cua.edu).

S.-H. Lin is with the Epson Palo Alto Laboratory, Palo Alto, CA 94304 USA (e-mail: shlin@epal.com).

H. Li is with the Department of Electrical Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: huaili@eng.umd.edu).

S.-Y. Kung is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kung@ee.princeton.edu).

Publisher Item Identifier S 1053-587X(98)08813-8.

represents a specific region of the knowledge space. This is indeed the case, and in particular, inspired by the principle of divide-and-conquer in applied statistics, probabilistic modular neural networks have become increasingly popular in the machine learning research [1], [4]–[7], [17], [36]. In this paper, we present a particular application of the probabilistic modular networks to the problem of mapping from databases. We describe a constructive criterion for designing the network architecture and the learning algorithm, both of which are governed by information theory [37]. The motivation of this work comes from following considerations. First, the database (available knowledge) and the network (learning capability) have been traditionally treated as two separate components in neural system design, where the relationship between them is not explicit [36]. It is desirable to have a network mapping a database, thus allowing an efficient information representation [25]. Second, since the complex cluster patterns and distributions intrinsically exhibited in a database are generally not transparent to the user, it will be difficult to interpret the output of system, to analyze the course of error, and to evaluate the process of performance [4]. A high-resolution divide-and-conquer architecture, i.e., hierarchy, may be required. Finally, in many practical applications, data mapping means either supervised (with objective of data classification) [2], unsupervised (with objective of data quantification) [12], [22], or the combined learning [5]. A flexible but unified scheme should be explored.

The quantitative mapping of a database may be decomposed into three distinctive learning tasks:

- 1) detection of the structure of each class model with locally mixture clusters;
- 2) estimation of the data distributions for each induced cluster inside each class;
- 3) classification of the data into classes that realizes the data memberships.

Although many previously proposed approaches have led to quite impressive results, several fundamental issues remain unresolved in the application domain. For example, the finite mixture model has very appealing properties to class distribution learning; the number of local clusters and the kernel shapes of cluster distributions are often assumed to be known, which is far from being realized in most applications [2], [9], [13], [17], [22]. The data mapping will be, in general, difficult to interpret since imposing a simple parametric model for the class may prevent the correct identification of the data structure [25] and the accurate estimation of the class boundaries [1], [26]. If the local models are to map the structure of the class

and the class boundaries, model selection must be taken into consideration on the goodness of fit [4], [7]. Furthermore, once the correct model is determined, we may formulate parameter learning as problem of maximum likelihood (ML) estimation [1], [2], [10]. The most popular algorithm in this domain is expectation–maximization (EM) algorithm [3], [19]. However, the EM algorithm has the reputation of being a slow algorithm since its batch training has a first-order convergence in which new information acquired in the expectation step is not used immediately [19], [21], [22]. In order to balance the tradeoff between efficiency and accuracy, on-line algorithms are proposed for large-scale sequential learning [3], [11] and are extended to supervised learning [6], [17]. The price to be paid is then a greatly increased memory requirements [20]. In addition, since data quantification (inside each class) and data classification (between the classes) may be the two independent objectives in applications, the optimality criteria for them are indeed different. However, the relationship between these two objectives, as well as how the error interferes each other, have not been fully understood [23], [26]. Moreover, empirical results indicate that many neural network classifiers, whose structure and learning rule were designed to directly approximate the class posterior probabilities, may be unnecessarily complex since the coupled training scheme has to adapt and update simultaneously both the class likelihood and the class prior probabilities [6], [25], [39].

The objective of this work is to propose a unified learning strategy for mapping a database: The main idea is to find, in a first place, a set of local mixture models that efficiently represent the data, together with a model selection procedure in which the optimal number and shape of the local clusters are found by the information-theoretic criteria. A partition of the data set into classes that indicate the membership of each data point may then be realized in a second phase, where the decision boundaries will be determined according to a supervised error-correction training. The major differences between our work and the previous work [1], [9], [15], [17], [20], [22], [25] are as follows.

- 1) We impose a model selection procedure to determine both the number and kernel shape of local clusters inside each class using information-theoretic criteria. This allows us to analyze how the result in model selection affects the performances of both data quantification and classification.
- 2) We apply a fully adaptive incremental algorithm to the unsupervised learning of the class distribution functions. It involves a *soft* classification of the data under the principle of least relative entropy, thus leading to an efficient and unbiased estimation.
- 3) We add a fine-tuning phase for learning decision likelihood boundaries using a reinforce or antireinforce supervision approach in which the class prior is adjusted in a separate phase.

This decoupled training scheme permits the use of high-capacity classifiers while maintaining a reasonable computational complexity for the further classification of the data into the classes. In addition, we have analyzed the pair-

wise relationships between quantification and classification, between *soft* and *hard* classification, and between unsupervised and supervised learning. The insights provide the guidance for the correct use of various methods in real-world applications.

The remainder of the paper proceeds as follows. Section II presents the problem formulation regarding the statistical modeling, unsupervised data quantification, and supervised data classification. This is followed by detailed description of the methods and algorithms that, in practice, appears to be the most complete of the approaches that we have studied. In Section III, three application examples in different domains are presented that illustrate the performance of the proposed techniques in various aspects. Major conclusions and discussions are summarized in the final section.

II. METHODS AND ALGORITHMS

A. Statistical Modeling

Recently, there has been considerable success in using finite mixture distributions and probabilistic modular networks for data quantification and classification [1], [3], [10], [17], [18], [34]. In order to validate the suitable stochastic models for data mapping with specified objectives, over the past few years, we have conducted an investigation into data statistics and derived several useful theorems [4], [12]. Assume that the data points x_i in a database come from M classes $\{\omega_1, \dots, \omega_r, \dots, \omega_M\}$, and each class contains K_r clusters $\{\theta_1, \dots, \theta_k, \dots, \theta_{K_r}\}$, where ω_r is the model parameter vector of class r , and θ_k is the kernel parameter vector of cluster k within class r . Further assume that in our training data set (which should be a representative subset of the whole database), each data point has a one-to-one correspondence to one of the classes denoted by its class label l_{ik}^* , defining a supervised learning task, but the true memberships of the data to the local clusters are unknown, defining an unsupervised learning task.

For the model of local class distribution, since the true cluster membership for each data point is unknown, we can treat cluster labels of the data as random variables denoted by l_{ik} [23]. By introducing a probability measure of a multinomial distribution with an unknown parameter π_k to reflect the distribution of the number of data points in each cluster, the relevant (sufficient) statistics are the conditional statistics for each cluster and the number of data points in each cluster. The class conditional probability measure for any data point inside the class r , i.e., the standard finite mixture distribution (SFMD), can be obtained by writing down the joint probability density of the x_i and l_{ik} and then, summing it over all possible outcomes of l_{ik} , as a sum of the general form

$$f(u|\omega_r) = \sum_{k=1}^{K_r} \pi_k g(u|\theta_k) \quad (1)$$

where $\pi_k = P(\theta_k|\omega_r)$ with a summation equal to one, and $g(u|\theta_k)$ is the kernel function of the local cluster distribution. Several observations are worth reiteration.

- 1) All data points in a class are identically distributed from a mixture distribution.

- 2) The SFMD model uses the probability measure of data memberships to the clusters in the formulation instead of realizing the true cluster label for each data point.
- 3) Since the calculation of the data histogram f_{x_r} from a class relies on the same mechanism as in (1), its values can be considered to be a sampled version of the true class distribution f_r^* .

For the model of global class distributions, we denote the Bayesian prior for each class by $P(\omega_r)$. Then, the sufficient statistics for mapping a database, i.e., the conditional finite mixture distribution (CFMD), is the pair of $\{P(\omega_r), f(u|\omega_r)\}$. According to the Bayes' rule, the posterior probability $P(\omega_r|x_i)$ given a particular observation x_i can be obtained by

$$P(\omega_r|x_i) = \frac{P(\omega_r)f(x_i|\omega_r)}{p(x_i)} \quad (2)$$

where $p(x_i) = \sum_{r=1}^M P(\omega_r)f(x_i|\omega_r)$. Again, several observations are worth reiteration:

- 1) In order to classify the data points into classes, (2) is a candidate as a discriminant function.
- 2) Since defining a supervised learning requires information of l_{ir}^* , the Bayesian prior $P(\omega_r)$ is an intrinsically known parameter and can be easily estimated by $P(\omega_r) = \sum_{i=1}^N l_{ir}^*/N$.
- 3) The only uncertainty comes from class likelihood function $f(u|\omega_r)$ that should be the key issue in the follow-on learning process.

For simplicity, in the following context we will omit class index r in our discussion when only single class distribution model is concerned and use θ to denote the parameter vector of regional parameter set $\{(\pi_k, \theta_k)\}$.

B. Data Quantification via Unsupervised Learning

The problem of data quantification addresses the combined estimation of regional parameters (π_k, θ_k) and detection of the structural parameter K_r and the kernel shape of $g(\cdot)$ in (1) based on the observations \mathbf{x}_r . One natural criterion used for learning the optimal parameter values is to minimize the distance between the SFMD, which is denoted by $f_r(u)$, and the class data histogram, which is denoted by $f_{x_r}(u)$ [3]. In this work, we use relative entropy (Kullback–Leibler distance), which was suggested by information theory [37], as the distance measure [for simplicity, we use $f_r(u)$ to denote $f(u|\omega_r)$ in our formulation] given by

$$D(f_{x_r}\|f_r) = \sum_u f_{x_r}(u) \log \frac{f_{x_r}(u)}{f(u|\omega_r)}. \quad (3)$$

Note that the new cost function overcomes the problems of using squared error by weighting errors more heavily when probabilities are near zero and one and diverging in the case of convergence at the wrong extreme [2], [11]. Furthermore, we have previously shown that when relative entropy is used as a distance measure, the distance minimization method is equivalent to the soft-split classification-based method under the criterion of maximum likelihood (ML) [12], [32]. The conclusion is summarized by the following theorem (see the proof in the Appendix):

Theorem 1: Consider a sequence of random variables x_1, \dots, x_{N_r} in \mathcal{R}^{N_r} . Assume that the sequence $\{x_i\}$ is independent and identically distributed (i.i.d.) by the distribution f_r . Then, the joint likelihood function $\mathcal{L}_r(\theta)$ is determined only by the histogram of data f_{x_r} and is given by

$$\mathcal{L}_r(\theta) = \exp(-N_r[H(f_{x_r}) + D(f_{x_r}\|f_r)]) \quad (4)$$

where H denotes the entropy with base e, and the maximization of joint likelihood function $\mathcal{L}_r(\theta)$ is equivalent to the minimization of relative entropy $D(f_{x_r}\|f_r)$.

Thus, data quantification is formulated as a distribution learning problem, and the actual optimality is achieved when this cost function reaches its minimum. However, statistical dependence between data points is one of some fundamental concerns in the problem formulation since the calculation of the data histogram assumes that all the data points are independent random variables. In order to validate the correct use of the (3) in data quantification, we prove the following theorem to show that the data histogram $f_{x_r}(u)$ converges to the true distribution $f_r^*(u)$ for all u with probability one as $N_r \rightarrow \infty$. Thus, when N_r is sufficiently large, minimization of the relative entropy between f_r and f_r^* can be well approximated by the minimization of the relative entropy between f_{x_r} and f_r . This fitting procedure can be practically implemented by maximizing the joint likelihood function under the independence approximation of the data (see proof in Appendix) [4].

Theorem 2: Consider a sequence of random variables x_1, \dots, x_{N_r} in \mathcal{R}^{N_r} . Assume that the sequence $\{x_i\}$ is asymptotically independent [40] and identically distributed by the finite normal mixture distribution f_r^* . For a closed convex set $E \subset \mathcal{F}_r$ and distribution $f_{x_r} \notin E$, let $f_r \in E$ be the distribution that achieves the minimum distance to f_{x_r} , i.e.,

$$f_r = \arg \min_{\mathcal{F}_r \in E} D(f_{x_r}\|f_r). \quad (5)$$

Then, when N_r approaches infinity, we have

$$\lim_{N_r \rightarrow \infty} D(f_r\|f_r^*) = 0 \quad (6)$$

with probability one, i.e., the estimated distribution of \mathbf{x}_r , given that f_r achieves the minimum of $D(f_{x_r}\|f_r)$ is close to f_r^* for large N_r .

Another important issue concerning unsupervised distribution learning is the detection of the structural parameters of the class distribution known as model selection [1]. The objective here is to propose a systematic strategy for determining the optimal number and kernel shape of local clusters when the prior knowledge is not available. The motivations are driven by various objectives and requirements in the real applications. For example, the prior knowledge on the true structure of a database is generally unknown, i.e., the number and the kernel shape of the local clusters are not available beforehand, and model selection is required in the data mapping procedure. This is indeed the case that is particularly critical in real clinical applications, where the structure of the disease patterns for a particular patient or for a particular type of cancer may be arbitrarily complex; therefore, correct identification and quantification of the information is very important [4],

[7]. Thus, it will be desirable to have a neural network structure that is adaptive in the sense that the number and kernel shape of local clusters are not fixed beforehand. One conventional approach for doing this is to use a sequence of hypothesis tests [3], [36]. The problem in this approach, however, is the subjective judgment in the selection of the threshold for different tests. Recently, there has been a great deal of interest in using information theoretic criteria, such as Akaike information criterion (AIC) [27], [34] and minimum description length (MDL) [28], [30], to solve this problem. The major thrust of this approach has been the formulation of a model fitting procedure in which an optimal model is selected from the several competing candidates such that the selected model best fits the observed data. For example, AIC will select the model that gives the minimum defined by

$$\text{AIC}(K_a) = -2 \log (\mathcal{L}(\hat{\theta}_{\text{ML}})) + 2K_a \quad (7)$$

where $\mathcal{L}(\hat{\theta}_{\text{ML}})$ is the likelihood of $\hat{\theta}_{\text{ML}}$, and K_a is the number of free adjustable parameters in the model. From a quite different point of view, MDL reformulates the problem explicitly as an information coding problem in which the best model fit is measured such that it assigns high probabilities to the observed data, while at the same time, the model itself is not too complex to describe [28]. A model is selected by minimizing the total description length defined by

$$\text{MDL}(K_a) = -\log (\mathcal{L}(\hat{\theta}_{\text{ML}})) + 0.5K_a \log N_r. \quad (8)$$

Note that, different from AIC, the penalty term in MDL takes into account the number of observations. However, the justifications for the optimality of these two criteria with respect to data quantification or classification are somewhat indirect and remain unresolved [3], [27], [32], and none of these approaches have directly addressed the problem of kernel shape learning [7].

In this work, we derive a new formulation of the information theoretic criterion [the minimum conditional bias/variance (MCBV) criterion] to solve model selection problem. Nevertheless, it was Akaike/Rissanen's work that was the inspirational source to this work, but some new interpretations are presented and justified with the information-theoretic means [32]. Our approach has a simple optimal appeal in that it selects a minimum conditional bias and variance model, i.e., if two models are about equally likely, MCBV selects the one whose parameters can be estimated with the smallest variance.

The new formulation is based on the fundamental argument that the value of the structural parameter can not be arbitrary or infinite because such an estimate might be said to have low "bias," but the price to be paid is high "variance" [31]. From Jaynes' principle, which is stated as "*the parameters in a model which determine the value of the maximum entropy should be assigned values which minimize the maximum entropy*" [29], let joint entropy of \mathbf{x} and $\hat{\theta}$ be $H(\mathbf{x}, \hat{\theta}) = H(\mathbf{x}|\hat{\theta}) + H(\hat{\theta})$, following the Bayes' law, a very neat interpretation states that the maximum of conditional entropy $H(\mathbf{x}|\hat{\theta})$ is precisely the negative of the logarithm of the likelihood function $\mathcal{L}(\mathbf{x}|\hat{\theta})$ corresponding to the entropy-maximizing distribution of \mathbf{x}

[28], [30]. Thus, we have

$$\max_{P_{\mathbf{x}}} H(\mathbf{x}|\hat{\theta}) = -\log (\mathcal{L}(\mathbf{x}|\hat{\theta}))|_{P_{\mathbf{x}}=\prod_{i=1}^{N_r} f_r(x_i)}. \quad (9)$$

Note that the uniformly randomization in the SFMD modeling corresponds to the maximum uncertainty [23], [37]. Furthermore, maximizing the entropy of the parameter estimates $H(\hat{\theta})$ results in

$$\max_{P_{\hat{\theta}}} H(\hat{\theta}) = \sum_{k=1}^{K_a} H(\hat{\theta}_k) \quad (10)$$

where when the variance of the parameter estimate is determined by the corresponding sample estimate, normal and independent distribution $P_{\hat{\theta}}$ gives the maximum entropy [37], [38].

Since the joint maximum entropy is a function of K_a and $\hat{\theta}$, by taking the advantage of the fact that model estimation is separable in components and structure, we define the MCBV criterion as

$$\text{MCBV}(K) = -\log (\mathcal{L}(\mathbf{x}|\hat{\theta}_{\text{ML}})) + \sum_{k=1}^{K_a} H(\hat{\theta}_{k\text{ML}}) \quad (11)$$

where $-\log (\mathcal{L}(\mathbf{x}|\hat{\theta}))$ is the conditional bias, and $\sum_{k=1}^{K_a} H(\hat{\theta}_k)$ is the conditional variance of the model. As both terms represent natural estimation errors about their true models and should be treated on an equal basis, a minimization leads to the characterization of the optimum estimation as

$$K_0 = \arg \left\{ \min_{1 \leq K \leq K_{\text{MAX}}} \text{MCBV}(K) \right\}. \quad (12)$$

That is, if the cost of model variance is defined as the entropy of parameter estimates, the cost of adding new parameters to the model must be balanced by the reduction they permit in the ideal code length for the reconstruction error. A practical MCBV formulation with code-length expression is further given by

$$\begin{aligned} \text{MCBV}(K) &= -\log (\mathcal{L}(\mathbf{x}|\hat{\theta}_{\text{ML}})) \\ &+ \sum_{k=1}^{K_a} \frac{1}{2} \log 2\pi e \text{Var}(\hat{\theta}_{k\text{ML}}). \end{aligned} \quad (13)$$

However, the calculation of $H(\hat{\theta}_{k\text{ML}})$ requires the true values of the model parameters that are to be estimated. It has been shown that if the number of observations exceeds the minimal value, the accuracy of the ML estimation tends quickly to the best possible accuracy determined by the Cramér–Rao lower bounds (CRLB's), as has been well studied theoretically in [1] and [38]. Thus, the CRLB's of the parameter estimates are used in the actual calculation representing the "conditional" bias and variance [33]. We have found that the new formulation for determining the value of K_0 exhibits a very good experimental performance that is consistent with both

AIC and MDL. It should be noted, however, that it is not the only plausible one; other criteria, such as cross validation techniques, may also be useful in this case.

The performance of model selection for two frequently used methods, i.e., the AIC and MDL, and the proposed criterion (MCBV) were first tested and compared in the simulation study. The computer-generated data was made up of four overlapping normal components. Each component represents one local cluster. The value for each component were set to a constant value, and the noise of normal distribution was then added to this simulation digital phantom. Three noise levels with different variance were set to keep the same signal-to-noise ratio (SNR), where SNR is defined as $10 \log_{10} (\Delta\mu)^2 / \sigma^2$, with $\Delta\mu$ being the mean difference between clusters, and σ^2 is the noise power. The original data for the simulation study are given in Fig. 1(a). The AIC, MDL, and MCBV curves, as functions of the number of local clusters K , are plotted in the same figure. According to the information-theoretic criteria, the minima of these curves indicate the correct number of the local cluster. From this experimental figure, it is clear that the number of local clusters suggested by these criteria are all correct. For larger noise level, the model selection based on the MCBV criterion provides a more differentiable result than the other two criteria. More application of the MCBV to the identification of real data structures will be presented in the next section.

As the counterpart for adaptive model selection, there are many numerical techniques to perform ML estimation of cluster parameters [3]. For example, EM algorithm first calculates the posterior Bayesian probabilities of the data through the observations and the current parameter estimates (E-step) and then updates parameter estimates using generalized mean ergodic theorems (M-step). The procedure cycles back and forth between these two steps. The successive iterations increase the likelihood of the model parameters. In order to obviate the need to store all the incoming observations and change the parameters immediately after each data point allowing for high data rates, we developed a probabilistic self-organizing mixture (PSOM) algorithm to solve the problem. This is a fully incremental and stochastic learning algorithm and is a generalized adaptive version of the similar algorithm we presented in [12]. The scheme provides winner-takes-in probability (Bayesian "soft") splits of the data, hence, allowing the data to contribute simultaneously to multiple clusters. For the sake of simplicity, we assume the kernel shape of local cluster to be a Gaussian with mean μ_k and variance σ_k^2 in the following derivation. By differentiating $D(f_{x_r} \| f_r)$ given in (3) (here, the index of cluster r is omitted) with respect to the unconstrained parameters μ_k and σ_k^2 , we obtain the standard gradient descent learning rule for the mean and variance parameter vectors ($k = 1, \dots, K$)

$$\mu_k^{(t+1)} = \mu_k^{(t)} + \frac{\lambda}{N} \sum_{i=1}^N (x_i - \mu_k^{(t)}) \frac{z_{ik}^{(t)}}{\sigma_k^{2(t)}} \quad (14)$$

$$\sigma_k^{2(t+1)} = \sigma_k^{2(t)} + \frac{\lambda}{N} \sum_{i=1}^N [(x_i - \mu_k^{(t)})^2 - \sigma_k^{2(t)}] \frac{z_{ik}^{(t)}}{2\sigma_k^{4(t)}} \quad (15)$$

where λ is the learning rate, and $z_{ik}^{(t)}$ is the posterior Bayesian probability defined by

$$z_{ik}^{(t)} = \frac{\pi_k^{(t)} g(x_i | \mu_k^{(t)}, \sigma_k^{2(t)})}{f(x_i | \theta)}. \quad (16)$$

By adopting a stochastic gradient descent scheme for minimizing $D(f_{x_r} \| f_r)$ [22], the corresponding on-line formulation is obtained by simply dropping the summation sign and updating the parameters after each stimulus presentation; this is equivalent to approximating, at each step, the sum on the right side of (14) and (15) with just one term randomly drawn from the N terms. Furthermore, we employ a learning rate adaptation to increase the rate of convergence through the adaptive stochastic gradient descent algorithm ($k = 1, \dots, K$) [35] as in

$$\mu_k^{(t+1)} = \mu_k^{(t)} + a(t)(x_{t+1} - \mu_k^{(t)}) z_{(t+1)k}^{(t)} \quad (17)$$

$$\sigma_k^{2(t+1)} = \sigma_k^{2(t)} + b(t)[(x_{t+1} - \mu_k^{(t)})^2 - \sigma_k^{2(t)}] z_{(t+1)k}^{(t)} \quad (18)$$

where the variance factors are incorporated into the learning rates, while the posterior Bayesian probabilities are kept, and $a(t)$ and $b(t)$ are introduced as the learning rates, two sequences converging to zero, ensuring unbiased estimates after convergence. The idea behind this update rule is motivated by the principle that every weight of a network should be given its own learning rate and that these learning rates should be allowed to vary over time [35]. Based on generalized mean ergodic theorem [37], updates can also be obtained for the constrained regularization parameters π_k in the SFMD model. For simplicity, given an asymptotically convergent sequence, the corresponding mean ergodic theorem, i.e., the recursive version of the sample mean calculation, should hold asymptotically [3]. From the M-step of EM algorithm, we define the interim estimate of π_k by

$$\pi_k^{(t+1)} = \frac{t}{t+1} \pi_k^{(t)} + \frac{1}{t+1} z_{(t+1)k}^{(t)}. \quad (19)$$

Hence, the updates given by (17)–(19) provide the incremental procedure for computing the SFMD component parameters. Their practical use, however, requires a strong mixing condition (data randomization) and a decaying annealing procedure (learning rate decay) [40]. These two steps are currently controlled by user-defined parameters that may not be optimized for a specific case. Therefore, algorithm initialization must be chosen carefully and appropriately [12], [32]. An overall convergence dynamics of the PSOM is similar to the competitive learning (CL) algorithm in that a solution is obtained by "resonating" between input data and an internal representation [36]. Such a mechanism can be considered to be a more realistic learning tool than the EM algorithm. In addition, the data distribution for each class can also be modeled by a finite generalized Gaussian mixture (FGGM) given by [34], where $g(x_i | \theta_k)$ is the generalized Gaussian kernel representing the k th local cluster's pdf, which is defined by

$$g(x_i | \theta_k) = \frac{\alpha \beta_k}{2\Gamma(1/\alpha)} \exp[-|\beta_k(x_i - \mu_k)|^\alpha], \quad \alpha > 0 \quad (20)$$

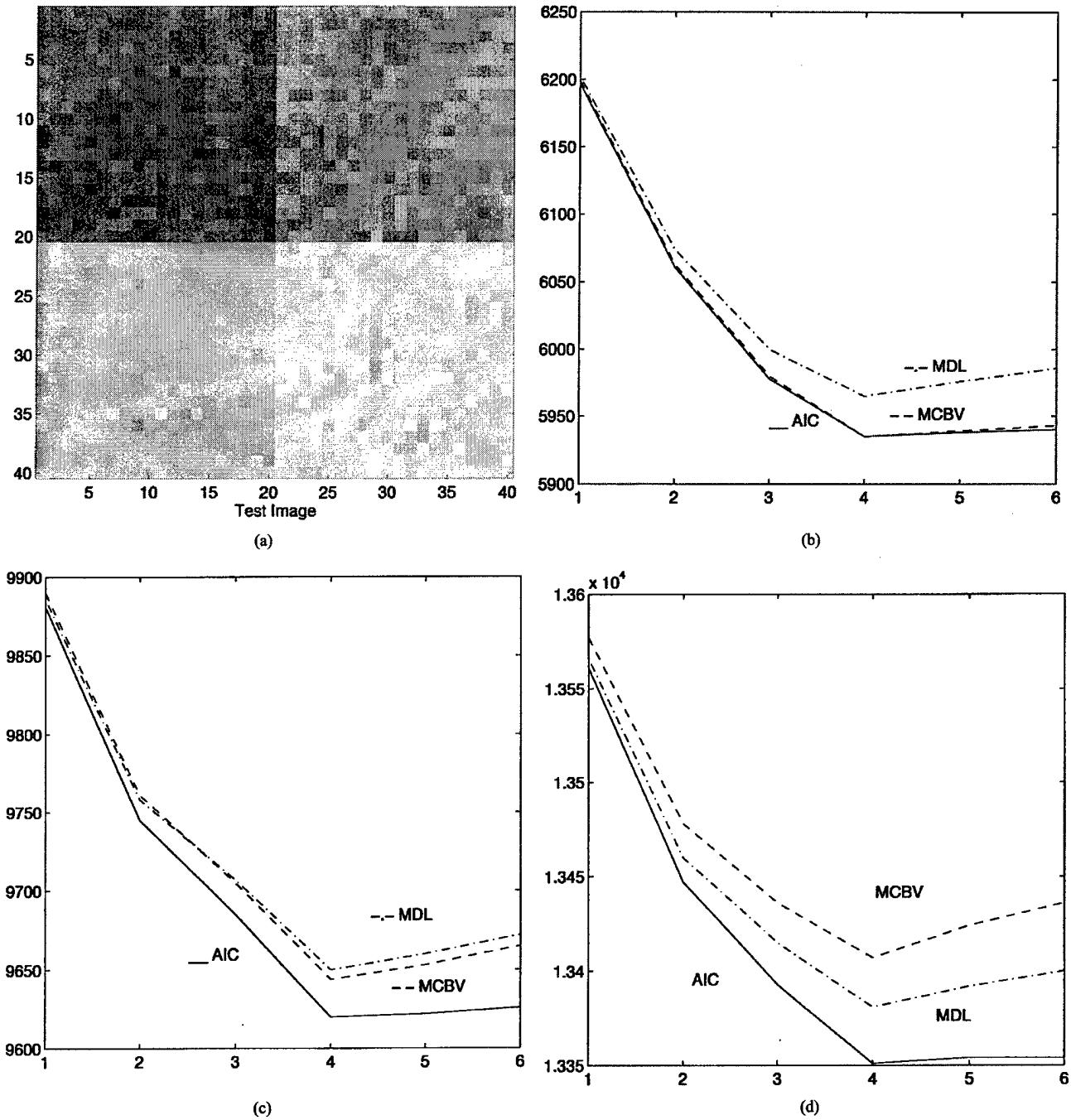


Fig. 1. Original test image ($K_0 = 4$, SNR = 10 dB) and the AIC/MDL/MCBV curves in model selection (left to right: $\sigma = 3, 30, 300$).

where

- μ_k mean;
- $\Gamma(\cdot)$ Gamma function;
- β_k parameter related to the variance σ_k by

$$\beta_k = \frac{1}{\sigma_k} \left[\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)} \right]^{1/2}. \quad (21)$$

It has been shown that when $\alpha = 2.0$, we have the Gaussian pdf; when $\alpha = 1.0$, we have the Laplacian pdf. When $\alpha \gg 1$, the distribution tends to a uniform pdf; when $\alpha < 1$, the pdf

becomes sharp. Therefore, the generalized Gaussian model is a suitable model for those data in which statistical properties are unknown, and the kernel shape can be controlled by selecting different α values.

C. Data Classification via Supervised Learning

The objective of data classification is to realize the class membership l_{ir} for all data points based on the observation x_i and the class statistics $\{P(\omega_r), f(u|\omega_r)\}$. It is well known that the optimal data classifier is the Bayes classifier since it can

achieve the minimum rate of classification error [38]. Measuring the average classification error by the mean squared error E , many previous researchers have shown that minimizing E by adjusting the parameters of class statistics is equivalent to directly approximating the posterior class probabilities when dealing with the two-class problem [2], [38]. In general, for the multiple class problem, the optimal Bayes classifier (minimum average error) classifies input patterns based on their posterior probabilities: Input x_i is classified to class ω_r if

$$P(\omega_r|x_i) > P(\omega_j|x_i) \quad (22)$$

for all $j \neq r$. It should be noted that in the formulation of classifier design, the optimal criterion used for the future data classification has been intuitively and directly applied to the learning of class statistics from the training data set.

Following this philosophy, great effort has been made in designing the network as an estimator of the posterior class probability [36]. By closely investigating the global class distribution modeling discussed in the previous section, we found that the classifier design for data classification can be dramatically simplified at the learning stage. Revisiting (2), since the class prior probability $P(\omega_r)$ is a known parameter when a supervised learning is applied, the posterior class probability $P(\omega_r|x_i)$ can be obtained without any further effort. Thus, by conditioning $P(\omega_r)$, the problem is formulated as a supervised classification learning of the class conditional likelihood density $f(u|\omega_r)$. It is very important that the learning process has been treated in a different way from the testing process while maintaining a consistency between the objective and the criterion. Moreover, when the ultimate goal of the learning is data classification, the question that may be asked is the following: Learning class likelihoods or decision boundaries? Since, in fact, only the decision boundaries are of the interests, the problem can be reformulated as the learning of the class boundaries (much more efficient) rather than the class likelihoods (generally time consuming). Thus, an efficient supervised algorithm to learn the class conditional likelihood densities called the "decision-based learning" [5] is adopted in this paper. The decision-based learning algorithm uses the *misclassified* data to adjust the density functions $f(u|\omega_r)$, which are initially obtained using the unsupervised learning scheme described previously so that the minimum classification error can be achieved. The algorithm is summarized as follows.

Define the r th-class discriminant function $\phi_r(x_i, \mathbf{w})$ to be $P(\omega_r)f(x_i|\omega_r)$. Given a set of training patterns $\mathbf{X} = \{x_i; i = 1, 2, \dots, M\}$. The set \mathbf{X} is further divided into the "positive training set" $\mathbf{X}^+ = \{x_i; x_i \in \omega_r, i = 1, 2, \dots, N\}$ and the "negative training set" $\mathbf{X}^- = \{x_i; x_i \notin \omega_r, i = N + 1, N + 2, \dots, M\}$. Define an energy function

$$E = \sum_{i=1}^M l(d(i)) \quad (23)$$

where

$$d(i) = \begin{cases} T - \phi_r(x_i, \mathbf{w}), & \text{if } x_i \in \mathbf{X}^+ \\ \phi_r(x_i, \mathbf{w}) - T, & \text{if } x_i \in \mathbf{X}^- \end{cases} \quad (24)$$

and where $T = \max_{\forall j \neq r}(\phi_j(x_i, \mathbf{W}))$. The *penalty function* l can be either a piecewise linear function

$$l(d) = \begin{cases} \zeta d, & \text{if } d \geq 0 \\ 0, & \text{if } d < 0 \end{cases} \quad (25)$$

where ζ is a positive constant or a sigmoidal function

$$l(d) = \frac{1}{1 + \exp^{-d\xi}}. \quad (26)$$

Notice that 1) energy function E is always large or equal to zero and 2) only misclassified training patterns contribute to the energy function. Therefore, the misclassification is minimized if E goes to the minimum.

The reinforced and antireinforced learning rules are used to update the network

Reinforced

$$\text{Learning: } \mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} + \eta l'(d(t)) \nabla \phi(\mathbf{x}(t), \mathbf{w})$$

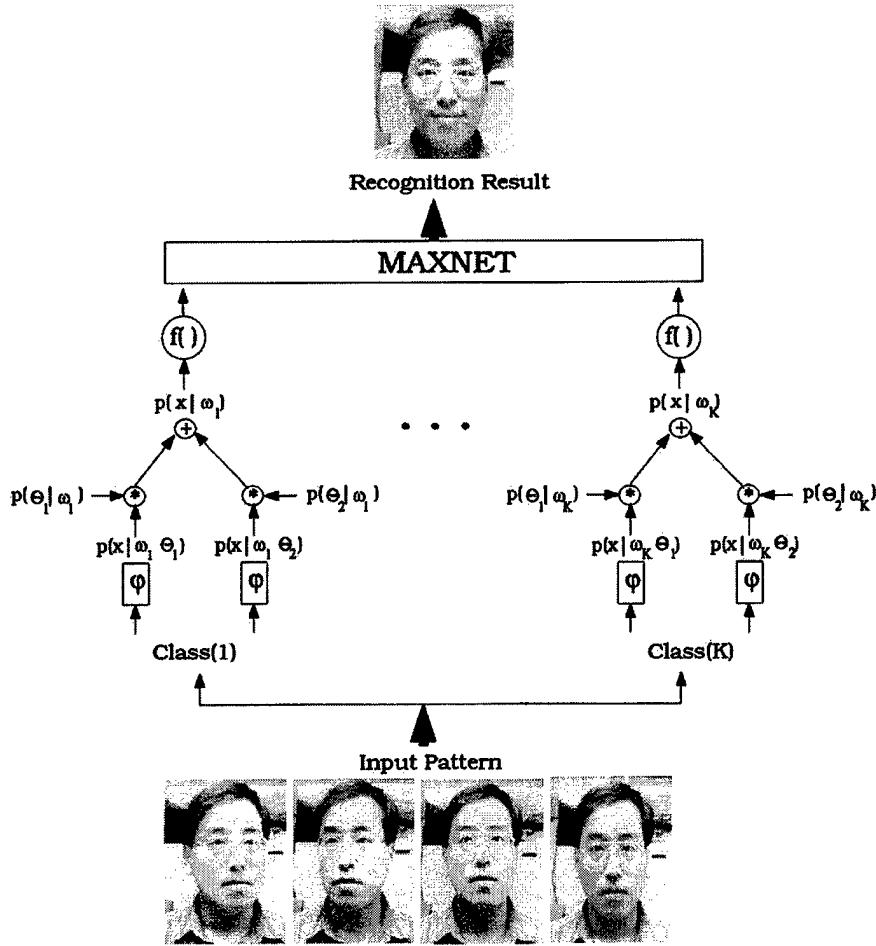
Antireinforced

$$\text{Learning: } \mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} - \eta l'(d(t)) \nabla \phi(\mathbf{x}(t), \mathbf{w}). \quad (27)$$

If the misclassified training pattern is from a positive training set, reinforced learning will be applied. If the training pattern belongs to the negative training set, we antireinforce the learning, i.e., pull the kernels away from the problematic regions.

A probabilistic decision-based neural network (PDBNN) [6] is a probabilistic modular network designed especially for data classification where a Bayesian decomposition of the learning process provides a unique opportunity to optimize the structure of training scheme [4], [6], [25]. Since the information about class population is, in general, physically uncorrelated with the conditional features about the individual class, a decoupled two-step training, in terms of both network structure and learning rule, makes much more sense than that in the conventional posterior-typed neural networks, i.e., the conditional likelihood of each class and the class Bayesian prior should be adjusted separately in the classification spaces. In theory, when the cost function in future classification is defined as the average Bayes' risk (with a discrete version of squared or mean squared classification error) [2], a sufficient measure field, which is determined by the average likelihood risk, can be applied in the supervised learning [6].

Thus, PDBNN divides its network resources into M different pieces, and each piece is designated to one data class only, i.e., the subnet outputs of the PDBNN are designed to model the likelihood functions (likelihood-typed network). As illustrated in Fig. 2, the structure of the PDBNN consists of several disjoint subnets and a winner-take-all network, where the class likelihood functions are first estimated from equally presented class samples, and the final decision boundaries are determined simply weighting the likelihood by the class populations. Clearly, by taking the advantage of availability of class prior in supervised training, the cost function can be redefined, the sample set can be reorganized, and both the network structure and learning process can be dramatically simplified [4]. For a M -classification problem, PDBNN contains M different class subnets, each of which represents one



(\textcircled{f}) : posterior type - normalization operator
likelihood type - log operator

Fig. 2. PDBNN network structure. Each class subnet is designated to recognize one class. All the network weightings are in probabilistic format.

data class in the database. Within each subnet, several neurons (or clusters) are applied in order to handle problems that have complicated decision boundaries. The outputs of class subnets are fed into a winner-take-all network. The winner-take-all network categorizes the input pattern to the data class whose subnet produces the highest output value. Recall our problem formulation in Section II-C; it becomes clear that each piece of the PDBNN is exactly a PSOM subnet. Thus, when the ultimate goal is data classification, all of the network parameters can now be initialized by the quantification (unsupervised learning) step before supervised training. This initialization, together with the fact that the number of hidden units in each PSOM is relatively small compared with that of the PDBNN, makes PDBNN achieve a faster convergence rate and, often, better classification accuracy.

The training scheme of the PDBNN is based on the so-called locally unsupervised globally supervised (LUGS) learning. There are two phases in this scheme: During the locally unsupervised (LU) phase, each subnet is trained individually,

and no mutual information across the classes may be utilized. Unsupervised algorithms such as the PSOM described in the previous section can be applied in this phase.

After the LU phase is completed, the training enters the globally supervised (GS) phase. In the GS phase, teacher information is introduced to reinforce or antireinforce the decision boundaries obtained during LU phase. There are three main aspects of this training phase.

- 1) *When to update:* A selective training scheme can be adopted, e.g., weight updating only when misclassification.
- 2) *What to update:* The learning rule is distributive and localized. It applies *reinforced learning* to the subnet corresponding to the correct class and *antireinforced learning* to the (unduly) winning subnet.
- 3) *How to update:* Adjust the boundary by updating the weight vector w either in the direction of the gradient of the discriminant function (i.e., reinforced learning) or the opposite of that direction (i.e., antireinforced learning).

Since only misclassified data points will be used for fine tuning of the decision boundaries, possible bias in the estimation of class distributions should be addressed. However, the key point we want to make is that this approach is very efficient, and although the global class description may be biased because of selective training, the decision boundaries will be more accurate. In fact, our intensive experiments indicate that only the data closed to the decision boundaries provide useful information in the boundary estimation. In particular, when the class distribution is formulated by a SFMD, the data far from the decision boundaries make little impact on the final classification results [6].

The discriminant functions in all clusters will be trained by the two-phase learning. A common model for the PDBNN to approximate the likelihood function is the mixture of Gaussians. The PDBNN designer can choose either hyperbasis function (HyperBF) or elliptical basis function (EBF) for the neurons to approximate full-rank or diagonal covariance matrices, respectively [6]. For the sake of simplicity, in this paper, we demonstrate the GS learning algorithm by using EBF only.

Suppose input pattern x_i is a D -dimensional vector $x_i = [x_i^1, x_i^2, \dots, x_i^D]^T$. Its EBF for cluster θ_k in class ω_r is

$$\psi(x_i, \omega_r, \theta_k) = -\frac{1}{2} \sum_{d=1}^D \beta_{rkd}(x_i^d - w_{rkd})^2 + C_{rk} \quad (28)$$

where $C_{rk} = -(D/2)(\ln 2\pi - \sum_{d=1}^D \ln \beta_{rkd})$. The initial values of the cluster parameters, i.e., β and w , can be obtained by PSOM. The discriminant function $\phi_r(x_i, \mathbf{w})$ for class r (see Section II-C) becomes

$$\phi_r(x_i, \mathbf{w}) = P(\omega_r) \sum_{k=1}^{K_r} \pi_k \exp(\psi(x_i, \omega_r, \theta_k)). \quad (29)$$

By applying reinforced and antireinforced learning rules in (29), β and w can further be updated. The gradient vectors for EBF at iteration j are computed as

$$\begin{aligned} \frac{\partial \phi_r(x_i, \mathbf{w})}{\partial w_{rkd}} \Big|_{\mathbf{w}=\mathbf{w}^{(j)}} &= h_{irk}^{(j)} \cdot \beta_{rkd}^{(j)}(x_i^d - w_{rkd}^{(j)}) \\ \frac{\partial \phi_r(x_i, \mathbf{w})}{\partial \beta_{rkd}} \Big|_{\mathbf{w}=\mathbf{w}^{(j)}} &= \frac{h_{irk}^{(j)}}{2} \left(\frac{1}{\beta_{rkd}^{(j)}} - (x_i^d - w_{rkd}^{(j)})^2 \right) \end{aligned} \quad (30)$$

$$h_{irk}^{(j)} = \frac{\pi_k^{(j)} \exp(\psi^{(j)}(x_i, \omega_r, \theta_k))}{\sum_l \pi_l^{(j)} \exp(\psi^{(j)}(x_i, \omega_r, \theta_l))}. \quad (31)$$

The cluster prior probabilities π_k can also be updated by

$$\pi_k^{(j+1)} = (1/N_r) \sum_{i=1}^{N_r} h_{irk}^{(j)}. \quad (32)$$

III. APPLICATION EXAMPLES AND DISCUSSIONS

A. Medical Image Quantification

In this section, we present the results using the information theoretic criteria to determine the appropriate number and/or kernel shape of tissue types (with a correspondence to the local

clusters) in the real MR brain images and digital mammograms as well as the results using the proposed quantification technique (e.g., the PSOM) to estimate the tissue quantities from these images. A fully automatic thresholding method, adaptive Lloyd-Max histogram quantization (ALMHQ) that we introduced recently in [12] is used to initialize the quantification, and the tissue parameters are then finalized by the PSOM. For the validation of the tissue quantification using the proposed algorithms, the global relative entropy (GRE) value is used as an objective measure to evaluate the accuracy of the data quantification, which is consistent with our problem formulation in Section II-B. The objective of the experiment is to illustrate the algorithm performance on real-world applications.

Fig. 3(a) and (b) show the original data consisting of two adjacent, T1-weighted images parallel to the anterior commissural-posterior commissural (AC-PC) line and the corresponding image histograms (c) and (d). This data were acquired with a General Electric (GE) Sigma 1.5 Tesla system. The imaging parameters are TR 35, TE 5, flip angle 45°, 1.5-mm effective slice thickness, 0 gap, 124 slices with in plane 192 × 256 matrix, and a 24-cm field of view. Since the skull, scalp, and fat in the original brain images do not contribute to the brain tissue, we edit the MR images to exclude nonbrain structures prior to tissue quantification [24]. Experience indicates that this procedure helps to achieve better quantification of brain tissues by delineation of the other tissue types that are not clinically interesting [9]. It can be clearly seen that the histograms have different shapes from slice to slice and that the tissue types are highly overlapped. This situation presents a great challenge to any computerized technique, even though it has been successful in the simulation study. In this study, in addition to the “gold standard” evaluation performed by neuroradiologists [8], we use the GRE value to reflect the quality of tissue quantification.

Based on pre-edited MR brain image, the procedure for quantifying the tissue types in one slice is summarized as follows.

- 1) For each value of K (number of tissue types), ML tissue quantification is performed by the PSOM algorithm.
- 2) Scan the values of $K = K_{\min}, \dots, K_{\max}$, and use the information-theoretic criteria to determine the suitable number of tissue types.
- 3) Select the result of tissue quantification corresponding to the value of K_0 determined in Step 2).
- 4) Evaluate the performance of tissue quantification in terms of the GRE value, convergence rate, and computational complexity.

In our experiment, since the number of tissue types is unknown, we first show that the number of tissue types varies from slice to slice. Let $K_{\min} = 2$ and $K_{\max} = 9$, and calculate $AIC(K)$, $MDL(K)$, and $MCBV(K)$ ($K = K_{\min}, \dots, K_{\max}$). We obtained the results shown in Fig. 4, which suggested that the two brain images contain six and eight tissue types, respectively. According to the model fitting procedure in designing the optimal structure of the modular networks we discussed before, the minima of these criteria also determines

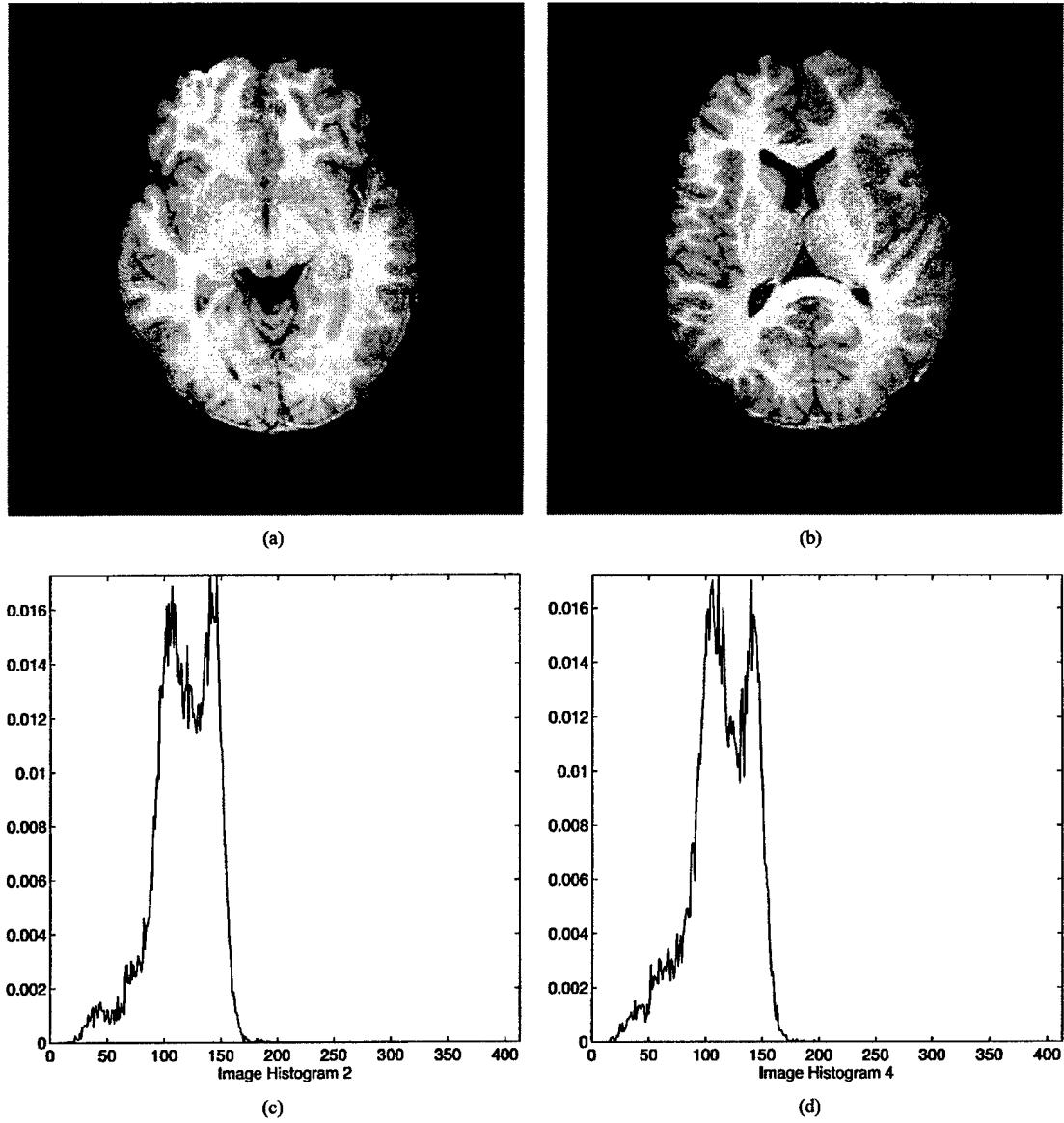


Fig. 3. Pure brain tissues extracted from (a) and (b) original MR images and (c) and (d) the corresponding histograms.

the most appropriate number of mixture components in the corresponding PSOM. These figures show that the overall performance of the three information-theoretic criteria is fairly consistent when applied to the real MR brain images. Our experience indicates, however, that AIC tends to overestimate while MDL tends to underestimate the number of tissue types, and MCBV provides the solution between those of AIC and MDL, which is believed to be more reasonable especially in terms of providing a balance between the bias and variance of the parameter estimates. As discussed in the literature, brain material is generally composed of three principal tissue types, i.e., WM, GM, CSF, and their pair-wise combinations known as the partial volume effect. Previous studies have proposed a six-tissue model representing the primary tissue types, and the mixture tissue types were defined as CSF-white (CW), CSF-gray (CG), and gray-white (GW). In this work, we also

consider the triple mixture tissue, which is defined by CSF-white-gray (CWG). More importantly, since the MRI scans clearly show the distinctive intensities at the local brain areas, the functional tissue types need to be considered. In particular, caudate nucleus and putamen are the two important local brain functional areas.

For each fixed K , the PSOM algorithm is iteratively used to quantify the different tissue types, where the learning is fully data-driven [12]. For slice 2, the results of final tissue quantification with $K_0 = 7, 8, 9$ are shown in Fig. 5. Corresponding to $K_0 = 8$, a GRE value of 0.02–0.04 nats in quantification is achieved. It was found that most of the variance parameters are different, which suggests that assuming the same variance for each tissue type with distinct image-intensity distribution may not be realistic. These quantified tissue types agreed with that of a physician's qualitative analysis results.

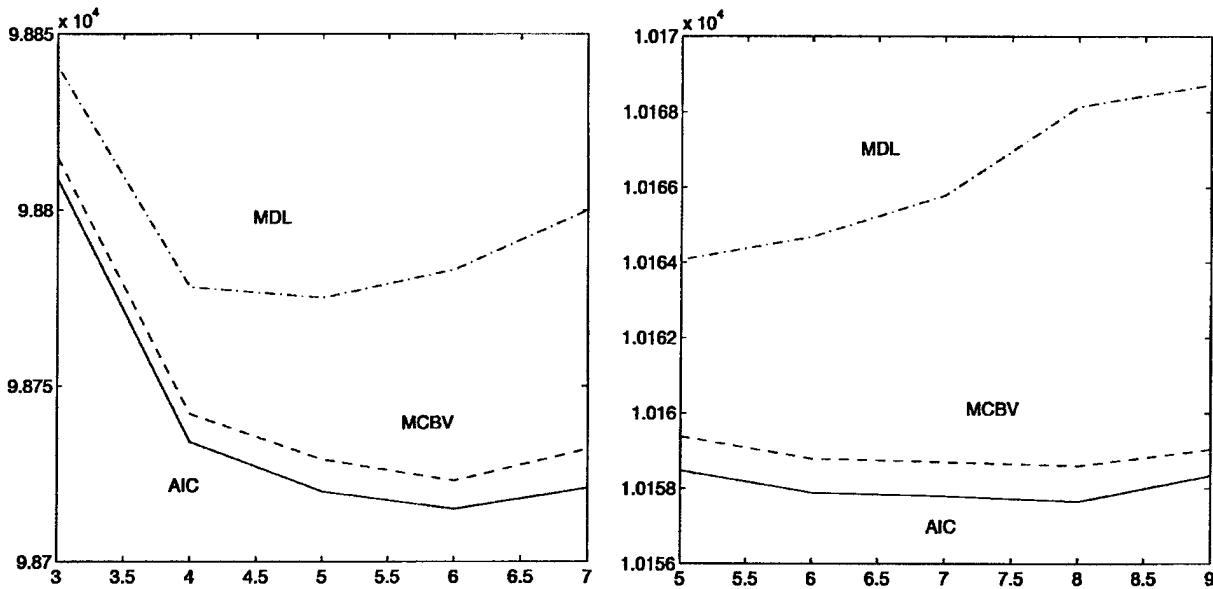


Fig. 4. Results of model selection for slice 1-2 ($K_0 = 6$ and 8, left to right).

We then present a comparison of the performance of PSOM with that of the EM [3], [19], [21] and the CL [6], [22] algorithms on MR brain tissue quantification. The task is to evaluate the computational accuracy and efficiency of the algorithm in the standard finite normal mixture distribution learning. To be able to make fair comparisons with the other two methods, we applied all the methods to the same example and used the GRE value between the image histogram and the estimated SFNM distribution as the goodness criterion to evaluate the quantification error. The left side of Fig. 6 shows learning curves of the PSOM and competitive learning (CL) averaged over five independent runs. As observed in the figure, PSOM outperforms CL learning by faster convergence rate and lower quantification error, where the final GRE value is about 0.04 nats. The right side of Fig. 6 presents the comparison of PSOM with that of the EM algorithm for 25 epochs. From the learning curves, again note that the PSOM algorithm shows superior estimation performance. The final quantification error is about 0.02 nats while preserving the faster convergence rate.

We have also applied the same procedure to the digital mammograms given in Fig. 7, where we show that if the number of cluster K is known, the kernel shape of local clusters will affect the accuracy of the histogram quantification for real mammographic images. Since, in this case, we do not assume a fixed kernel shape, FGGM is used, and three information criteria (AIC, MDL, and MBVC) were used to determine both the number and kernel shape of the regions in the digital mammograms. Twenty real mammograms with masses were chosen as testing images. The selected mammograms were digitized with an image resolution of $100 \mu\text{m} \times 100 \mu\text{m}$ per pixel by the laser film digitizer (Model Lumiscan 150). The image sizes are $1792 \times 2560 \times 12$ b/pixel. We found that, although with different α , all three criteria achieved minimum when $K = 8$. It indicates that these information criteria are relatively insensitive to the change of α , as also claimed

in [34]. With this observation, we can further decouple the relation between K and α and choose the appropriate value of one while fixing the value of another. It is interesting to note that the result of model selection here is very consistent with the conclusion in some previous studies: according to the work in [41], the most appropriate region number (K) is eight for most digital mammograms. We fixed $K = 8$, and changed the values of α for estimating the FGGM model parameters using the PSOM/EM algorithm. The GRE value between the histogram and the estimated FGGM distribution is used as a measure of the estimation bias. We found that GRE achieved a minimum value when $\alpha = 3.0$ as shown in Fig. 8. Compared with the conventional finite normal mixture model ($\alpha = 2.0$), which has been mostly chosen by many previous researchers, this experiment indicates that the FGGM model provides more freedom, thus allowing its correct uses to the situation when the true statistical properties of the digital mammograms are not available.

B. Face Recognition Experiment

A PDBNN-based face recognition system [6] is being developed under a collaboration between Siemens Corporate Research, Princeton, NJ, and Princeton University, Princeton, NJ. The total system diagram is depicted in Fig. 9. All four main modules—face detector, eye localizer, feature extractor, and face recognizer—are implemented on a SUN Sparc10 workstation. An RS-170 format camera with 16 mm, F1.6 lens is used to acquire image sequences. The S1V digitizer board digitizes the incoming image stream into 640×480 8-bit gray-scale images and stores them into the frame buffer. The image acquisition rate is on the order of 4–6 frames/s. The acquired images are then down sized to 320×240 for the following processing.

As shown in Fig. 9, the processing modules are executed sequentially. A module will be activated only when the in-

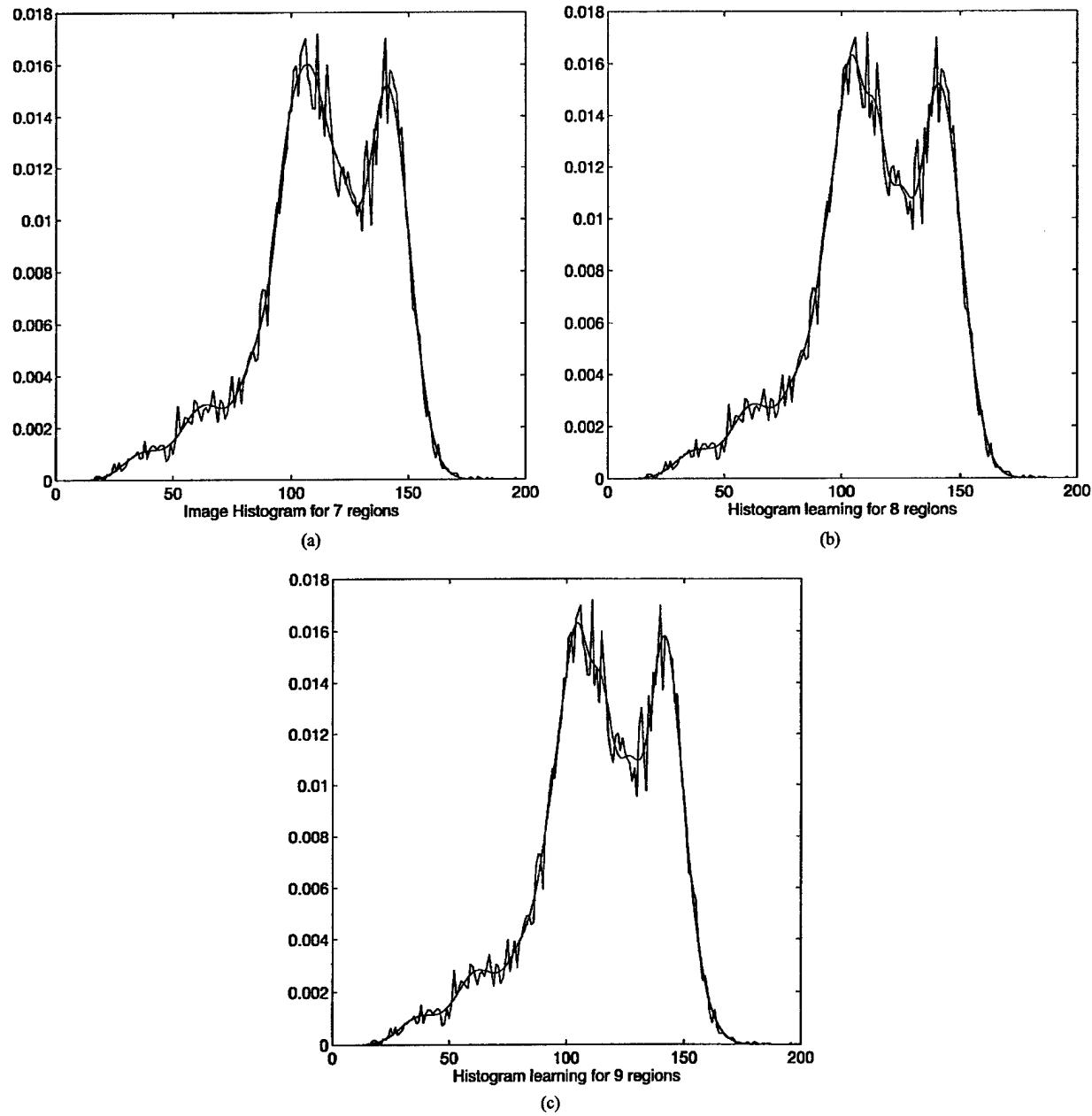


Fig. 5. Histogram learning for slice 2. (a) $K_0 = 7$. (b) $K_0 = 8$. (c) $K_0 = 9$.

coming pattern passes the preceding module (with an agreeable confidence). After a scene is obtained by the image acquisition system, a quick detection algorithm based on binary template matching is applied to detect the presence of a proper sized moving object. A PDBNN face detector is then activated to determine whether there is a human face. If positive, a PDBNN eye localizer is activated to locate both eyes. A subimage ($\approx 140 \times 100$) corresponding to the face region will then be extracted. Finally, the feature vector is fed into a PDBNN face recognizer for recognition and subsequent verification.

The system built on the proposed one has been demonstrated to be applicable under reasonable variations of orientation and/or lighting and with the possibility of eyeglasses. This method has been shown to be very robust against large varia-

tion of face features, eye shapes, and cluttered background [6]. The algorithm takes only 200 ms to find human faces in an image with 320×240 pixels on a SUN Sparc10 workstation. For a facial image with 320×240 pixels, the algorithm takes 500 ms to locate two eyes. In the face recognition stage, the computation time is linearly proportional to the number of persons in the database. For a 200-person database, it takes less than 100 ms to recognize a face. Furthermore, because of the inherent parallel and distributed processing nature of PDBNN, the technique can be easily implemented via specialized hardware for real-time performance.

We conduct an experiment on the face database from the Olivetti Research Laboratory, Cambridge, U.K. (the ORL database). There are ten different images of 40 different

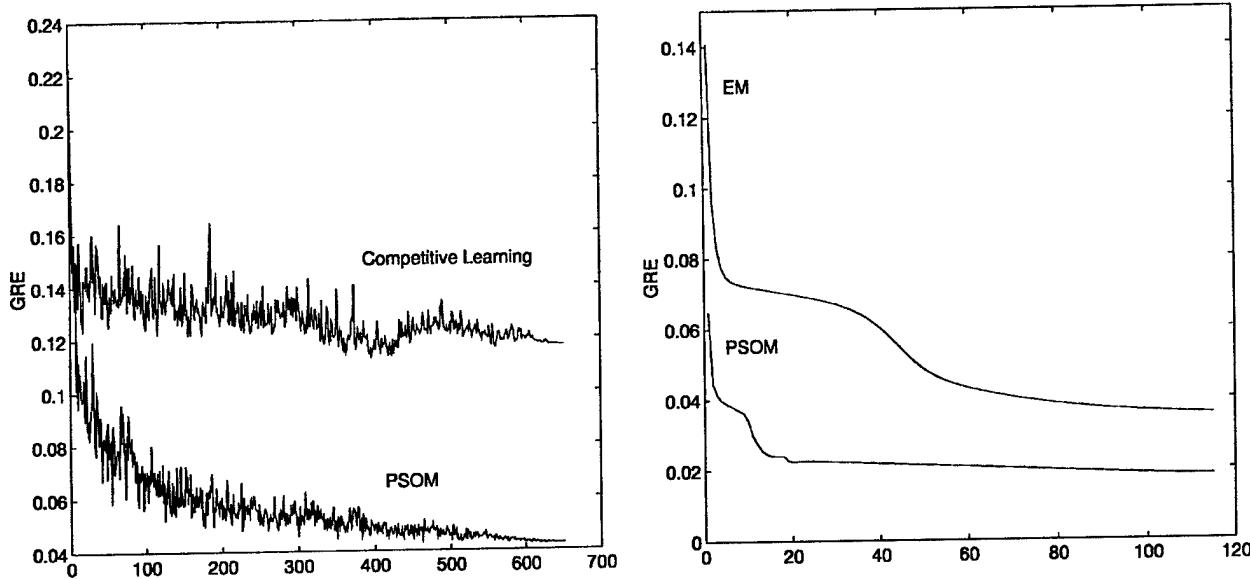


Fig. 6. Comparison of the learning curves of (left) PSOM and CL and (right) EM.

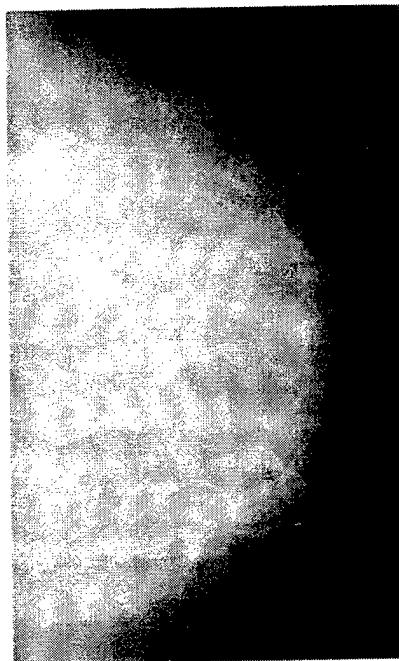


Fig. 7. Typical image of an original digital mammogram.

people. There are variations in facial expression (open/close eyes, smiling/nonsmiling), facial details (glasses/no glasses), scale (up to 10%), and orientation (up to 20°). A HMM-based approach is applied to this database and achieves 13% error rate [13]. The popular eigenface algorithm [16] reports the error rate around 10% [13], [14]. In [15], a pseudo 2-D HMM method is used and achieves 5% at the expense of long computation time (4 m/pattern on Sun Sparc II). In [14], Lawrence *et al.* use the same training and test set size as Samaria did as well as a combined neural network (self organizing map and convolutional neural network) to do the

TABLE I
PERFORMANCE OF DIFFERENT FACE RECOGNIZERS ON THE ORL DATABASE.
PART OF THIS TABLE IS ADAPTED FROM S. LAWRENCE *et al.*,
“FACE RECOGNITION: A CONVOLUTIONAL NEURAL NETWORK
APPROACH,” TECHNICAL REPORT, NEC RESEARCH INSTITUTE, 1995

System	Error rate	Classification time	Training Time
PDBNN	4%	< 0.1 seconds	20 minutes
SOM + CN	3.8%	< 0.5 seconds	4 hours
Pseudo 2D-HMM	5%	240 seconds	n/a
Eigenface	10%	n/a	n/a
HMM	13%	n/a	n/a

recognition. This scheme spent 4 hr to train the network and less than 1 s to recognize one facial image. The error rate for the ORL database is 3.8%. Our PDBNN-based system reaches similar performance (4%) but has much faster training and recognition speed (20 m for training and less than 0.1 s for recognition). Both approaches run on SGI Indy. Table I summarizes the performance numbers on the ORL database.

We have also applied the PDBNN method to the so-called “M + 1 classes” problem in which the pattern under testing could be either from one of the M classes or from some other unknown class (the “unknown” class or the “intruder” class). Note that the unknown class probability is often very hard to estimate, and for some applications, it is almost impossible to obtain enough training samples for the unknown class (for example, in the face recognition problem, the unknown class includes the faces all over the world). In our experiment, PDBNN uses a different decision rule from that of the “M class” problem: Pattern x_i belongs to class r if both of the following conditions are true: a) $\phi(\omega_r, x_i) > \phi(\omega_j, x_i), \forall j \neq r$, and b) $\phi(\omega_r, x_i) > T$, T is a threshold obtained by decision-based learning. Otherwise, pattern x_i belongs to the unknown

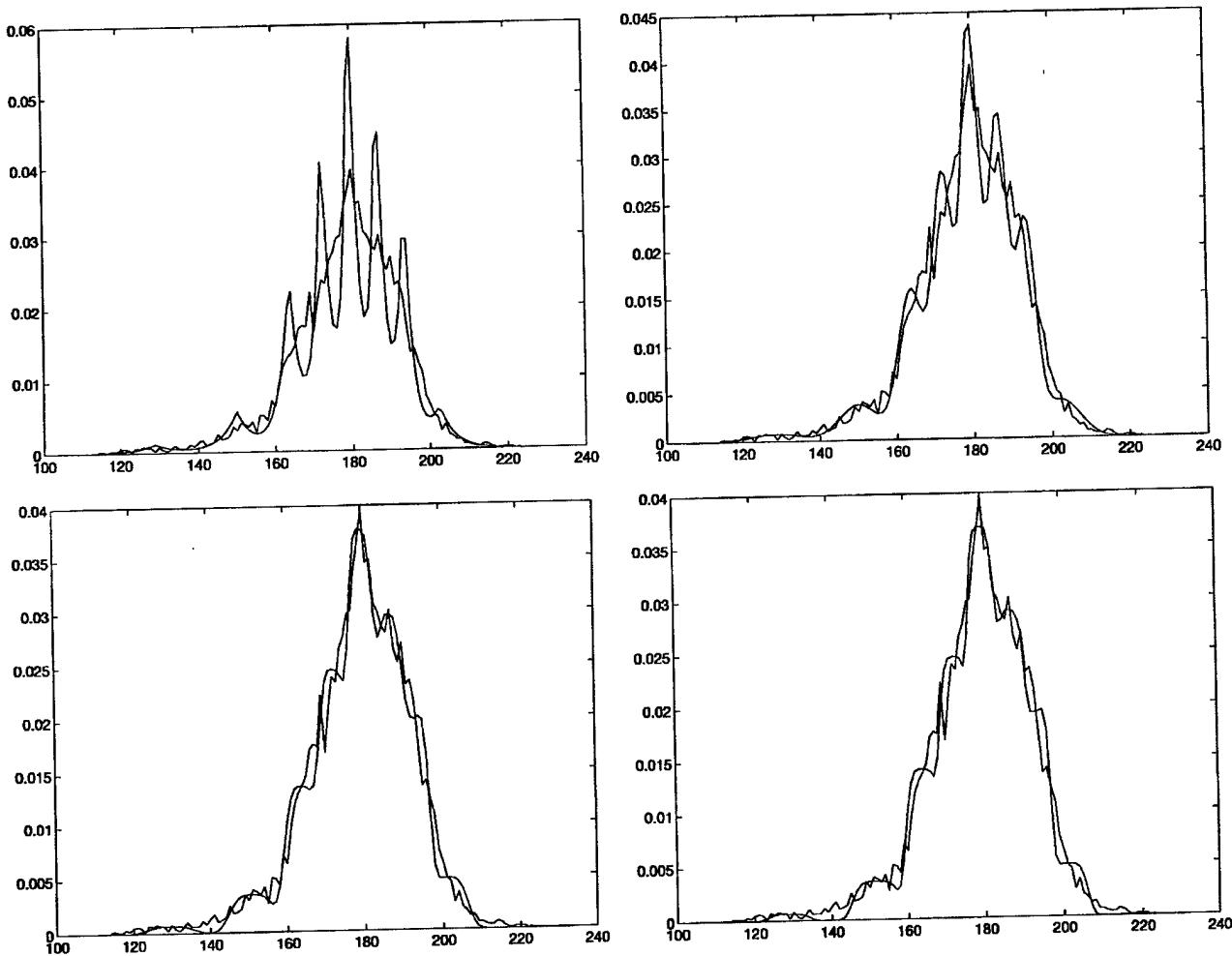
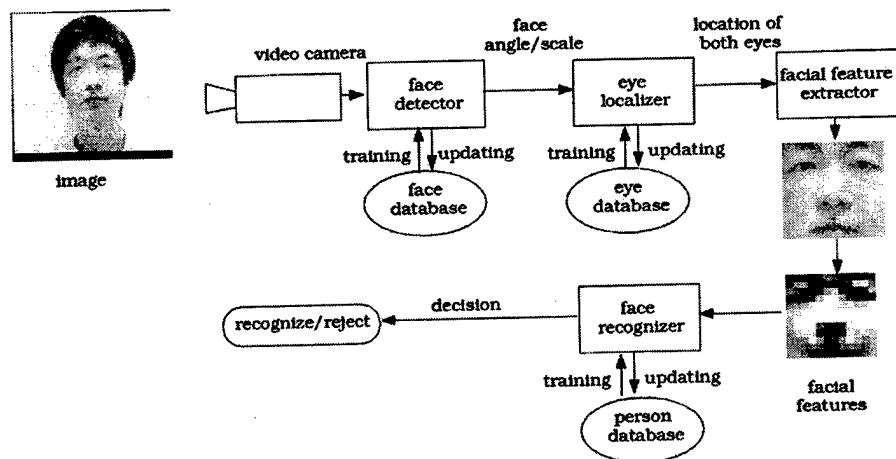
Fig. 8. Comparison of mammogram histogram learning with different kernel shapes ($K_0 = 8$).

Fig. 9. System configuration of the face recognition system. Face recognition system acquires images from video camera. Face detector determines if there are faces inside images. Eye localizer indicates the exact positions of both eyes. It then passes their coordinates to facial feature extractor to extract low-resolution facial features as input of face recognizer.

class. We observed consistent and significant improvement in classification results, comparing pure Bayesian decision and the PDBNN approach (e.g., recognition rate from 70–90%) contributed by the fine-tuning process [6]. The following

example further shows the effect of the fine-tuning process: For 100-person face recognition, we have 500 training patterns/person and 20 test patterns/person. After the LU phase, we obtained a training accuracy of 89.2% (44 608/50 000) and

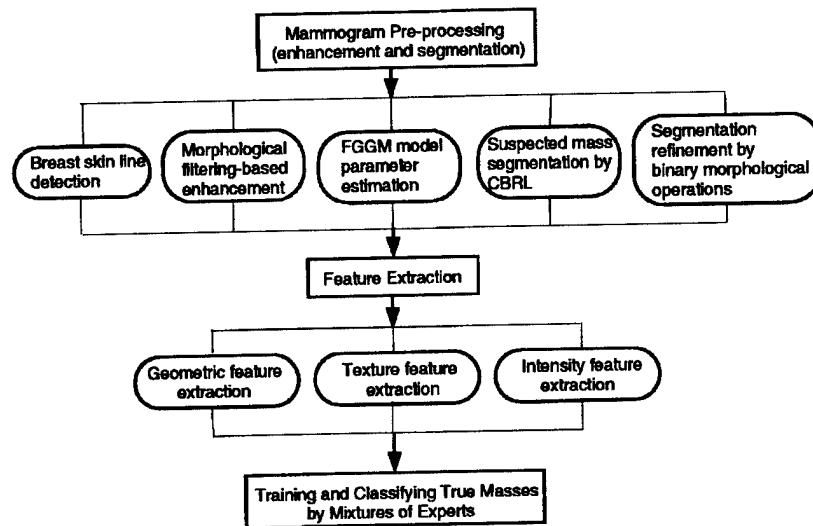


Fig. 10. Flow diagram of mass detection in digital mammograms.

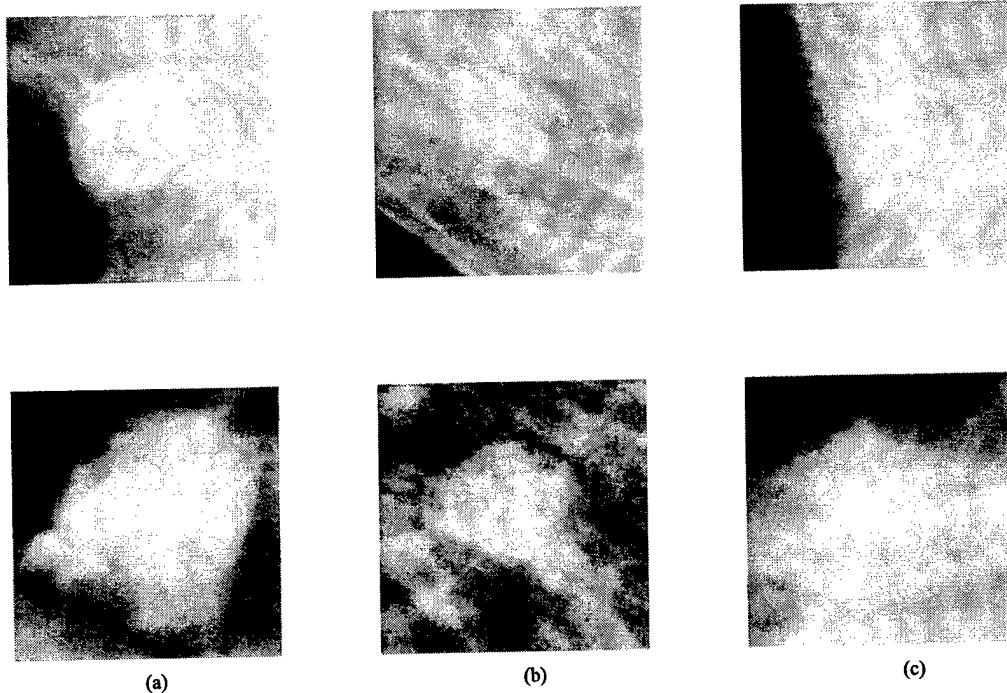


Fig. 11. Typical mass appearances in mammograms. (a) Well-defined masses. (b) Ill-defined masses. (c) Spiculated masses.

a test accuracy of 71.5% (1430/2000). After the GS phase, we improved the performance to a training accuracy of 98.9% (49 495/50 000) and a test accuracy of 96.2% (1924/2000). Nevertheless, when we have the luxury of knowing the object probability model in advance, the fine-tuning process may not be necessary. It is reasonable to acknowledge that the face recognition result from our experiment is valid since the ORL database is a widely used public database like the FERET database. With a comparison with the recognition rate of the eigenface method on an early FERET database (smaller size), we found that the performance of the proposed method is comparable and/or superior to the eigenface approach.

C. Featured Database Analysis

As we have discussed in Sections I and II, model selection is the first and a very important learning task in mapping a database, and the objective of the procedure is to determine both the number and kernel shape of local clusters in each class. The inaccuracy in model selection will affect the performances of both data quantification and classification. Using the proposed learning scheme, the structure of the probabilistic modular networks will be optimized following the model selection and PSOM [7], [32]. When all the class distributions are learned accurately, further data classification

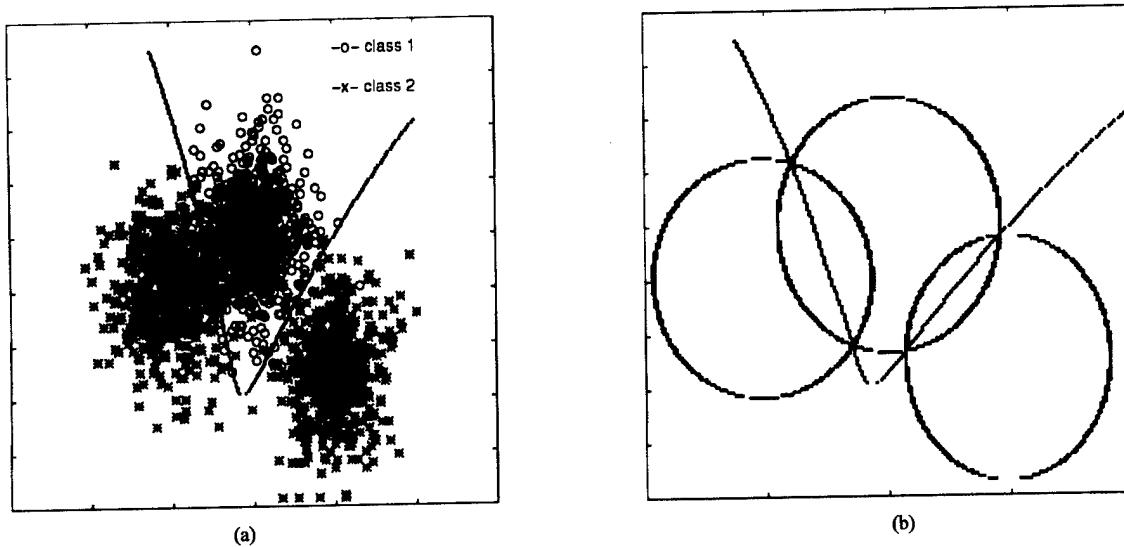


Fig. 12. Two-dimensional feature space in classification example 1 where “o” denotes true mass cases; “*” denotes false mass cases. (a) Class 2 contains two clusters. (b) Decision boundary learning with four cross points.

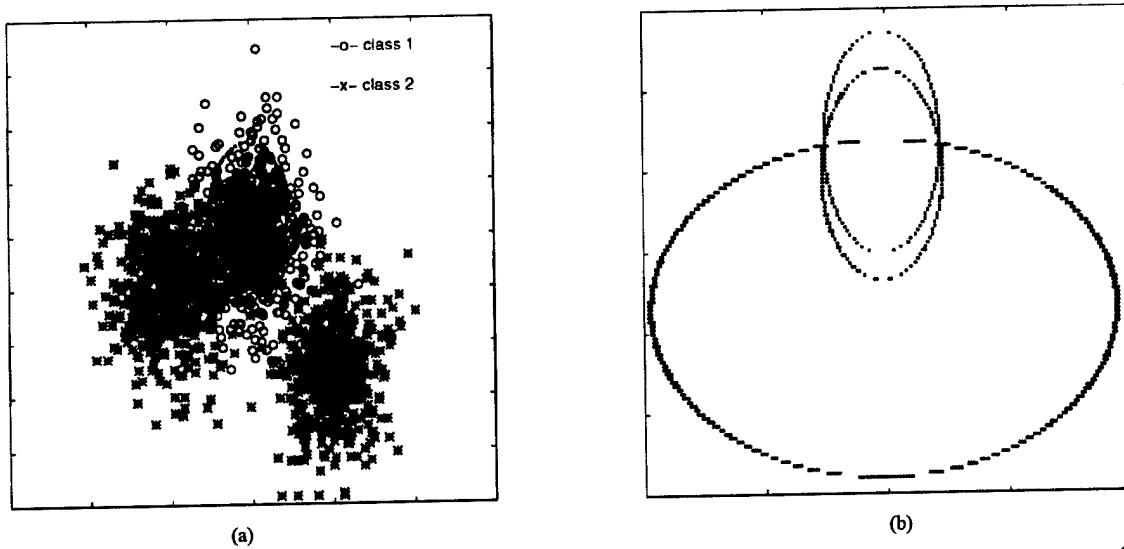


Fig. 13. Two-dimensional feature space in classification example 2, where “o” denotes true mass cases, and “*” denotes false mass cases. (a) Class 2 contains one cluster. (b) Decision boundary learning with two cross points.

will be achieved simply following Bayesian rule [38]. In this subsection, these objectives and the related conclusions are further illustrated by two examples in the computed-aided diagnosis (CAD) for breast cancer detection [7]. The objective is to detect masses in digital mammography since masses are the important signs leading to early breast cancer [7]. For the purpose of improving the performance of CAD for detection of early breast cancer in mammography, a crucial step in any strategic solution is to quantitatively analyze the featured database (with the cases of normal and cancer tissues), i.e., to create a map of the feature distributions regarding the disease patterns [4], [7]. Since the featured database in CAD is constructed from the preprocessed suspected regions, model selection is very important in providing useful diagnostic suggestions. Furthermore, based on the feedback after all possible lesions are detected and their features are quantified,

database quality and learning capability of the CAD system design can also be analyzed by the model selection, comparing different feature extraction and database construction schemes [4]. The framework of the proposed method for mass detection is illustrated in Fig. 10.

Some typical mass cluster appearances on mammograms are displayed in Fig. 11. With a preprocessing step, all suspected mass regions, as well as some normal dense tissues with brighter intensities, are located. The latter should be eliminated from the true masses through feature discrimination. On the clinical site, masses are evaluated based on the location, density, size, shape, margins, and the presence of associated calcifications.

In the first example, we show that the inappropriate determination of the number of clusters inside each class will affect the performance of data classification. Since a classi-

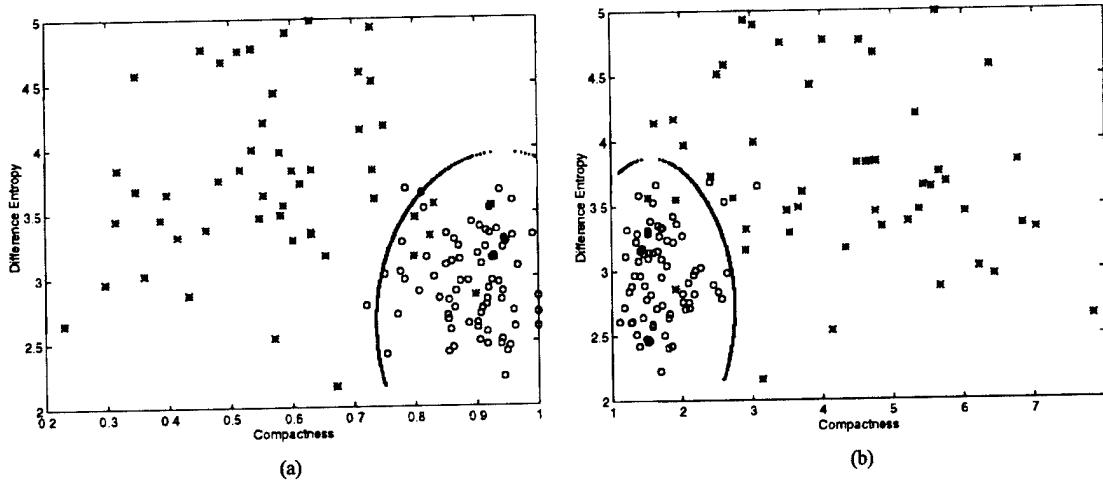


Fig. 14. Classification results. —○— denotes true mass cases; —*— denotes false mass cases. (a) Classification using compactness definition 1. (b) Classification using compactness definition 2.

fication based on feature space is commonly used in many pattern analysis applications, including mammographic mass detection, typical intensity, geometric, and texture features are extracted and investigated from the segmented regions. These features usually possess clinical significance and are widely used in most CAD systems. A detailed description of feature extraction can be found in [7]. Suppose we extract two major features that characterize the two targeted classes (mass and nonmass), as it shown in Fig. 12. In this example, class 1 contains one cluster, and class 2 contains two clusters. The 2-D histogram pairs of these features extracted from true and false mass regions are investigated, and the features that can better separate the true and false mass regions are selected for further study. In this study, area, compactness (circularity), and difference entropy were found to have better discrimination and reliability properties. Therefore, we chose them to perform the classification.

Two PDBNN-like modular networks are trained to classify these two classes. The classification results are shown in Figs. 12 and 13. The result in Fig. 12 is with the right cluster number in Class 2. The result in Fig. 13 is with the wrong cluster number in Class 2. In this simple experiment, it is clearly shown that comparing the results in Fig. 12 with those in Fig. 13, the classification boundary with the right cluster number may be much more accurate than that with the heuristically determined cluster number since the decision boundary between classes 1 and 2 will be determined by four cross points in the first case, whereas in the second case, the decision boundary will be determined by only two cross points. From this example, we can show that the error of data classification is controlled by the accuracy in estimating the decision boundaries between classes, and the quality of the boundary estimates is indeed dependent on both the bias and variance of the class likelihood estimates. It can be seen that the bias may be lower in case 1 than in case 2, but the variance will be higher in case 1 than case 2. A similar example is the curve fitting from noisy data [31].

In the second example, we use the proposed classifier to distinguish true masses from false masses based on the

features extracted from the suspected regions. The objective is to reduce the number of suspicious regions and identify the true masses. We selected 150 mammograms from the mammographic database. Each mammogram contained at least one mass case of varying size and location. The areas of suspicious masses were identified by an expert radiologist based on visual criteria and biopsy-proven results. We selected 50 mammograms with biopsy-proven masses from the data set for training. The mammogram set used for testing contained 46 single-view mammograms: 23 normal cases and 23 with biopsy-proven masses. The feature vector contained two features: compactness and difference entropy. According to our investigation, these two features have the better separation (discrimination) between the true and false mass classes. These features are also not correlated with each other. According to our experience, the values of compactness with definition 1 are more reliable than those of compactness with [7, Def. 2]. A training feature vector set was constructed from 50 true-mass ROI's and 50 false-mass ROI's. The training set was used to train two modular probabilistic decision-based neural networks separately. Fig. 14(a) shows the classification of two classes with compactness definition 1. Fig. 14(b) shows the classification of two classes with compactness definition 2.

In our evaluation study, six to 15 suspected masses per mammogram were detected and required further evaluation. The receiver operating characteristic (ROC) method is used to evaluate the detection performance of our method [38]. In the ROC analysis, the distribution of the positive and negative cases can be represented by certain probability distributions. When the two distributions overlap on the decision axis, a cutoff point can be made at an arbitrary decision threshold. The corresponding true-positive fraction (TPF) versus false-positive fraction (FPF) for each threshold can be drawn on a plane. By indicating several points on the plot, curve fitting can be employed to construct an ROC curve. The area under the curve, which is referred to as A_z , can be used as a performance index of the system. In general, the higher the A_z , the better the performance. In addition, two other indexes [sensitivity (TPF) and specificity (1-FPF)] are usually used to

evaluate the system performance on the specified point of the ROC curve. In this study, a computer program (LABROC) is employed for the evaluation analysis. We found that the proposed classifier can reduce the number of suspicious masses with a sensitivity of 84% at a specificity of 82% (1.6 false positive findings per mammogram) based on the database containing 46 mammograms (23 of them have biopsy-proven masses). In conclusion, when compared with the conventional neural networks, the probabilistic modular networks can lead to more efficient learning and provide better understanding in the analysis of the distribution patterns of multiple features extracted from the suspicious masses.

IV. CONCLUSIONS AND DISCUSSIONS

We have presented a strategy for mapping a database by probabilistic modular networks and information-theoretic criteria. Local class distribution is modeled by a standard finite mixture. Information-theoretic criteria are applied to detect the number and shape of local clusters, thus allowing the corresponding neural network to adaptively evolve its structure to the best representation of the local data. The PSOM algorithm is used to quantify the parameters of the local clusters, leading to an ML estimation. The decision boundaries in the data classification are then fine tuned by a global supervised learning. The results obtained by using the simulated data and the real databases demonstrate the promise and effectiveness of the proposed technique.

Our main contribution is the complete proposal of a de-tripled learning strategy for the determination of both modular and components of the network. In this approach, the network structures (in terms of which statistical model is more suitable) are justified in a first step and followed by a soft classification of the data (in terms of each data point supports all local clusters simultaneously). The associated probabilistic class labels are then realized in a third step as the competitive learning of this induced hard classification task. To summarize, the results of the experiments we have performed indicate the plausibility of this approach for database mapping and show that it can be applied to practical and clinical problems such as those encountered in face recognition and computer-aided diagnosis.

Model selection for the first time explicitly incorporates the bias/variance dilemma in finite data training, and when tested with synthetic and actual data, the results show that the number of hidden nodes should be adjusted for both data quantification and data classification, thus leading to a unified framework. At issue is how the model selection affects the estimation error and how the error in the estimation of class likelihoods further affects the classification error when the estimates are used in a classification rule. However, none of previously developed methods has directly addressed a goal of minimizing classification errors, which is a central objective of data classification. It is necessary, therefore, to develop methods that are more directly related to the minimization of classification errors. On the other hand, many previous researchers have shown that one of the most fundamental problems in detection and estimation is the bias/variance

dilemma [25], [26], [30], [31]. It has been reported that the bias and variance components of the estimation error combine to influence classification in a very different way than with squared error on the likelihoods themselves [1], [25], [26]. Their results also suggested that the bias and variance components may not be treated in an equal base for further improving the classifier's performance [26], and a minimum entropy approach was proposed for model selection aiming at maximizing the class separability [1]. However, these methods may be found to be problematic when the accuracy of both data quantification and classification is required.

Further comparison of the data quantification to the data classification calls for the following pair-wise relationships in the learning paradigm (supervised and unsupervised) and in the implementing scheme (*soft* and *hard*). In fact, when data quantification is the objective, unsupervised learning is preferred where only a *soft* classification of the data is required [23]. More precisely, since maximum likelihood is the criterion, local cluster parameters can be learned without *hard* data classification [1], [12], [22], [24]. If this unsupervised process involves a *hard* classification of a sample into the cluster for which the posterior probability is maximum, such as in the *k*-means algorithm [22], the quantities obtained by the sample averages after data classification may not be consistent with the previous quantification result since a perfect classification may not be possible when the distributions of local clusters are highly overlapping [23]. The quantification result, in general, will be biased. On the other hand, in order to perform data classification for the testing set where the objective is to minimize the average Bayes' risk, supervision is needed at a first place and can be realized by simply dividing the training set (e.g., a subset of the testing set) into the groups for the estimation of each local class likelihood (e.g., unsupervised learning of local clusters), whereas the global class Bayesian prior can be picked up immediately as the by-product of the dividing process. In this research, we deal with data quantification for local clusters and data classification between classes as two separate problems and use different optimality criteria. However, it is worth reiteration that in order to efficiently determine the decision boundaries between classes in data classification, supervised and unsupervised training may be jointly performed.

APPENDIX COLLECTED PROOFS OF THE THEOREMS

Proof of Theorem 1: Since the multiplication over i in joint likelihood is not affected by the data order, we regroup them in an increasing order of the gray levels u_i such that $u_1 < u_2, \dots, < u_L$. Hence, we write

$$\mathcal{L}_r(\theta) = \prod_{i=1}^{N_r} f_r(x_i) = \prod_{l=1}^L \left(\prod_{x_i=u_l} f_r(x_i) \right). \quad (33)$$

By the definition of data histogram (i.e., the type) in [37], the number of data with gray level u_l equals $N_r f_{x_r}(u_l)$; thus, we

have

$$\begin{aligned}
 \mathcal{L}_r(\theta) &= \prod_{l=1}^L f_r(u_l)^{N_r f_{\mathbf{x}_r}(u_l)} \\
 &= \prod_{l=1}^L \exp(N_r f_{\mathbf{x}_r}(u_l) \log f_r(u_l)) \\
 &= \prod_{l=1}^L \exp(N_r [f_{\mathbf{x}_r}(u_l) \log f_r(u_l) \\
 &\quad - f_{\mathbf{x}_r}(u_l) \log f_{\mathbf{x}_r}(u_l) \\
 &\quad + f_{\mathbf{x}_r}(u_l) \log f_{\mathbf{x}_r}(u_l)])] \\
 &= \exp\left(-N_r \sum_{l=1}^L \left[f_{\mathbf{x}_r}(u_l) \log \frac{1}{f_{\mathbf{x}_r}(u_l)}\right.\right. \\
 &\quad \left.\left.+ f_{\mathbf{x}_r}(u_l) \log \frac{f_{\mathbf{x}_r}(u_l)}{f_r(u_l)}\right]\right) \\
 &= \exp(-N_r [H(f_{\mathbf{x}_r}) + D(f_{\mathbf{x}_r} \| f_r)]). \quad \square
 \end{aligned}$$

Proof of Theorem 2: For each data value u_l , we apply indicator function $I(\cdot, u_l)$ to data sequence \mathbf{x}_r . By the definition of histogram, we have the relationship between the histogram $f_{\mathbf{x}_r}(u_l)$ and the sample average of the indicator functions $I(x_i, u_l)$. Since sequence \mathbf{x} is asymptotically independent and identically distributed by the finite normal mixture distribution, they are ergodic processes. In addition, since the indicator function is a deterministic measurable function, by the Birkhoff-Khinchin theorem [40]

$$\Pr\left(\lim_{N_r \rightarrow \infty} \frac{1}{N_r} \sum_{i=1}^{N_r} I(x_i, u_l) = E[I(x_i, u_l)]\right) = 1. \quad (34)$$

Since, by the fundamental theorem of expectation, we have

$$E[I(x_i, u_l)] = \sum_u I(x_i = u, u_l) f_r^*(u) = f_r^*(u_l) \quad (35)$$

we can substitute (3) and (9) into (8) to obtain

$$\Pr\left(\lim_{N_r \rightarrow \infty} f_{\mathbf{x}_r}(u_l) = f_r^*(u_l)\right) = 1$$

which implies that the distance of $D(f_{\mathbf{x}_r} \| f_r^*)$ goes to 0 as $N_r \rightarrow \infty$.

We now show that the estimated distribution f_r is close to f_r^* for large N_r in relative entropy. By the "Pythagorean" theorem ([37, Th. 12.6.1])

$$D(f_{\mathbf{x}_r} \| f_r) + D(f_r \| f_r^*) \leq D(f_{\mathbf{x}_r} \| f_r^*) \quad (36)$$

which in turn implies that

$$D(f_r \| f_r^*) \leq D(f_{\mathbf{x}_r} \| f_r^*) \quad (37)$$

since $D(f_{\mathbf{x}_r} \| f_r) \geq 0$. Note that the relative entropy $D(f_{\mathbf{x}_r} \| f_r^*)$ behaves like the square of the Euclidean distance [37]. From the conditions given by the theorem, the angle between the distances $D(f_{\mathbf{x}_r} \| f_r)$ and $D(f_r \| f_r^*)$ must be

obtuse, which implies (36). Consequently, since $D(f_{\mathbf{x}_r} \| f_r^*) \rightarrow 0$, it follows that

$$\lim_{N_r \rightarrow \infty} D(f_r \| f_r^*) = 0 \quad (38)$$

as $N_r \rightarrow \infty$ with probability one. \square

REFERENCES

- [1] L. Perlovsky and M. McManus, "Maximum likelihood neural networks for sensor fusion and adaptive classification," *Neural Networks*, vol. 4, pp. 89–102, 1991.
- [2] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 1361–1364.
- [3] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.
- [4] Y. Wang, "Database mapping by mixture of experts in computer-aided diagnosis," Tech. Rep., Georgetown Univ. Med. Cent., Washington, DC, July 1996.
- [5] S. Y. Kung and J. S. Taur, "Decision-based neural networks with signal/image classification applications," *IEEE Trans. Neural Networks*, vol. 1, pp. 170–181, Jan. 1995.
- [6] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE Trans. Neural Networks, Special Issue on Artificial Neural Networks and Pattern Recognition*, vol. 8, Jan. 1997.
- [7] H. Li et al., "Detection of masses on mammograms using advanced segmentation techniques and an HMOE classifier," in *Proc. 3rd Int. Workshop Digital Mammography*, Chicago, IL, June 1996, pp. 397–400.
- [8] P. Santiago and H. D. Gage, "Quantification of MR brain images by mixture density and partial volume modeling," *IEEE Trans. Med. Imag.*, vol. 12, pp. 566–574, Sept. 1993.
- [9] A. J. Worth and D. N. Kennedy, "Segmentation of magnetic resonance brain images using analog constraint satisfaction neural networks," *Inform. Process. Med. Imag.*, pp. 225–243, 1993.
- [10] D. P. Helmbold, R. E. Schapire, Y. Singer, and M. K. Warmuth, "A comparison of new and old algorithms for a mixture estimation problem," Tech. Rep., Univ. Calif., Santa Cruz and AT&T Lab., 1996.
- [11] E. Weinstein, M. Feder, and A. V. Oppenheim, "Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1652–1654, Sept. 1990.
- [12] Y. Wang and T. Adali, "Efficient learning of finite normal mixtures for image quantification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, 1996, pp. 3422–3425.
- [13] F. S. Samaria and A. C. Harter, "Parameterization of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vision*, Sarasota, FL, 1994.
- [14] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural network approach," Tech. Rep., NEC Res. Inst., 1995.
- [15] F. S. Samaria, "Face recognition using hidden markov model," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 1994.
- [16] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, pp. 71–86, 1991.
- [17] M. I. Jordan and R. A. Jacobs, "Hierarchical mixture of experts and the EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.
- [18] C. E. Priebe, "Adaptive mixtures," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 910–912, 1994.
- [19] R. A. Redner and N. M. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, pp. 195–239, 1984.
- [20] R. M. Neal and G. E. Hinton, "A new view of the EM algorithm that justifies incremental and other variants," *Biometrika*, 1993.
- [21] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixture," Tech. Rep., Artif. Intell. Lab., Mass. Inst. Technol, Cambridge, Jan. 1995.
- [22] J. L. Marroquin and F. Girosi, "Some extensions of the K-means algorithm for image segmentation and pattern classification," Tech. Rep., Artif. Intell. Lab., Mass. Inst. Technol, Cambridge, Jan. 1993.
- [23] D. M. Titterington, "Comments on 'application of the conditional population-mixture model to image segmentation,'" *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 656–658, Sept. 1984.
- [24] Y. Wang and T. Adali, "Probabilistic neural networks for parameter quantification in medical image analysis," *Biomed. Eng. Recent Development*, 1994.

- [25] J. L. Marroquin, "Measure fields for function approximation," *IEEE Trans. Neural Networks*, vol. 6, pp. 1081-1090, May 1995.
- [26] J. H. Friedman, "On bias, variance, 0/1-loss, and the curse-of-dimensionality," Tech. Rep., Stanford Univ., Stanford, CA, 1996.
- [27] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, Dec. 1974.
- [28] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Stat.*, vol. 11, no. 2, 1983.
- [29] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 108, no. 2, pp. 620-630/171-190, May 1957.
- [30] J. Rissanen, "Minimax entropy estimation of models for vector processes," *Syst. Identification: Advances Case Studies*, pp. 97-119, 1987.
- [31] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, pp. 1-52, 1992.
- [32] Y. Wang, "Image quantification and the minimum conditional bias/variance criterion," in *Proc. 30th Conf. Inform. Sci. Syst.*, Princeton, NJ, Mar. 20-22, 1996, pp. 1061-1064.
- [33] L. I. Perlovsky, "Cramer-Rao bounds for the estimation of normal mixtures," *Pattern Recognit. Lett.*, vol. 10, pp. 141-148, 1989.
- [34] J. Zhang and J. M. Modestino, "A model-fitting approach to cluster validation with application to stochastic model-based image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 1009-1017, Oct. 1990.
- [35] R. A. Jacobs, "Increased rates of convergence through learning rate adaptation," *Neural Networks*, vol. 1, pp. 295-307, 1988.
- [36] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: MacMillan, 1994.
- [37] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [38] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1988.
- [39] A. S. Pandya and R. B. Macy, *Pattern Recognition with Neural Networks in C++*. Boca Raton, FL: CRC, 1996.
- [40] R. Gray and L. Davisson, *Random Processes—A Mathematical Approach for Engineers*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
- [41] M. J. Bianchi, A. Rios, and M. Kabuka, "An algorithm for detection of masses, skin contours, and enhancement of microcalcifications in mammograms," in *Proc. Comput.-Assisted Radiol.*, Winston-Salem, NC, June 1994, pp. 57-64.



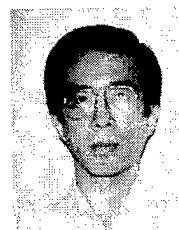
Yue Wang received the B.S. and M.S. degrees from Shanghai Jiao Tong University, Shanghai, China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of Maryland, Baltimore County, Baltimore, in 1995, all in electrical engineering.

He is currently with the Department of Electrical Engineering and Computer Science, Catholic University of America, Washington, DC, as an Assistant Professor. He is also affiliated with the Department of Radiology, Georgetown University School of Medicine, Washington, DC, as an Adjunct Assistant Professor. His research interests include image analysis, medical imaging, information visualization, database mapping, volumetric display, visual explanation, and their applications in biomedicine and multimedia informatics. Dr. Wang is the recipient of a 1998 U.S. Army Medical Research Command Career Development Award.



Shang-Hung Lin received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1991. He received the M.S. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 1994 and 1996, respectively.

He is currently with Epson Palo Alto Laboratory, Palo Alto, CA. His primary research interests include neural networks, pattern recognition, computer vision, and image processing.



Huai Li received the B.S. degree in engineering physics from Tsinghua University, Beijing, China, in 1985, the M.Med. degree in biomedical engineering from Beijing Medical University in 1988, and the M.S. and Ph.D. degrees in electrical engineering from University of Maryland, College Park, in 1995 and 1997, respectively.

Since 1997, he has been with the multimedia team at Odyssey Technologies, Inc., Jessup, MD, and is currently a Member of the Technical Staff, working on image/video processing and telecommunication.

His research interests include medical image analysis, image processing, video coding, and telecommunication.



Sun-Yuan Kung (F'88) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

In 1974, he was an Associate Engineer of Amdahl Corporation, Sunnyvale, CA. From 1977 to 1987, he was a Professor of Electrical Engineering Systems, University of Southern California, Los Angeles. Since 1987, he has been a Professor of Electrical Engineering at Princeton University, Princeton, NJ. He has authored more than 300 technical publications, including three books: *VLSI Array Processors* (Englewood Cliffs, NJ: Prentice-Hall, 1988) (with Russian and Chinese translations), *Digital Neural Networks* (Englewood Cliffs, NJ: Prentice-Hall, 1993), and *Principal Component Neural Networks* (New York: Wiley, 1996).

Dr. Kung received the 1992 IEEE Signal Processing Society's Technical Achievement Award for his contributions on parallel processing and neural network algorithms for signal processing. Since 1990, he has served as Editor-in-Chief of the *Journal of VLSI Signal Processing*. Recently, he served as a General Chair of the 1997 IEEE Workshop on Multimedia Signal Processing at Princeton University. He was appointed an IEEE-SP Distinguished Lecturer in 1994. He received the 1996 IEEE Signal Processing Society's Best Paper Award.

Computerized Radiographic Mass Detection—Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation

Huai Li, Yue Wang, K. J. Ray Liu*, Shih-Chung B. Lo, and Matthew T. Freedman

Abstract—This paper presents a statistical model supported approach for enhanced segmentation and extraction of suspicious mass areas from mammographic images. With an appropriate statistical description of various discriminative characteristics of both true and false candidates from the localized areas, an improved mass detection may be achieved in computer-assisted diagnosis (CAD). In this study, one type of morphological operation is derived to enhance disease patterns of suspected masses by cleaning up unrelated background clutters, and a model-based image segmentation is performed to localize the suspected mass areas using stochastic relaxation labeling scheme. We discuss the importance of model selection when a finite generalized Gaussian mixture is employed, and use the information theoretic criteria to determine the optimal model structure and parameters. Examples are presented to show the effectiveness of the proposed methods on mass lesion enhancement and segmentation when applied to mammographical images. Experimental results demonstrate that the proposed method achieves a very satisfactory performance as a preprocessing procedure for mass detection in CAD.

Index Terms—Finite mixture, image enhancement, image segmentation, information criterion, morphological filtering, relaxation labeling.

I. INTRODUCTION

IN RECENT years, several computer-assisted diagnosis (CAD) schemes for mass detection and classification have been developed [1]–[13]. Though it may be difficult to compare the relative performance of these methods, because the reported performance strongly depends on the degree of subtlety of masses in the selected database, accurate selection

Manuscript received February 3, 1997; revised January 9, 2001. This work was supported in part by the Department of Defense under Grants DAMD17-98-1-8045 and DAMD17-96-1-6254 through a subcontract from University of Michigan, Ann Arbor, and by the National Science Foundation (NSF) under NYI Award MIP-9457397. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. Giger. Asterisk indicates corresponding author.

H. Li is with the Electrical Engineering Department and Institute for Systems Research, University of Maryland at College Park, College Park, MD 20742 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Y. Wang is with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

*K. J. Ray Liu is with the Electrical Engineering Department and Institute for Systems Research University of Maryland at College Park, College Park, MD 20742 USA (e-mail: kjrliu@eng.umd.edu).

S.-C. B. Lo and M. T. Freedman are with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Publisher Item Identifier S 0278-0062(01)02831-2.

of suspected masses is considered a critical and first step due to the variability of normal breast tissue and the lower contrast and ill-defined margins of masses [3], [6], and since no subtle masses should be missed before any further analysis.

A number of image processing techniques have been proposed to perform suspicious mass site selection. Kobatake *et al.* [1] proposed using a iris filter to detect tumors as suspicious regions with very weak contrast to their background. Sameti *et al.* [7] used fuzzy sets to partition the mammographic image data. Lau and Yin *et al.* independently proposed using bilateral-subtraction to determine possible mass locations [9], [13]. Some other investigators proposed using pixel-based feature segmentation of spiculated masses [4], [8]. Kegelmeyer has reported promising results for detecting spiculated tumors based on local edge characteristics and Laws texture features [8]. Karssemeijer *et al.* [4] proposed to identify stellate distortions by using the orientation map of line-like structures. Recently, Petrick *et al.* [6] proposed a two-stage adaptive density-weighted contrast enhancement filtering technique along with edge detection and morphological feature classification for automatic segmentation of potential masses. Kupinski and Giger [3] presented a radial gradient index-based algorithm and a probabilistic algorithm for seeded lesion segmentation.

Nevertheless, to our best knowledge, few work has been dedicated to improve the task of lesion site selection although it is indeed a very crucial step in CAD. Especially, few studies have used and justified model-based image processing techniques for unsupervised lesion site selection [11]. Zwiggelaar *et al.* developed a statistical model to describe and detect the abnormal pattern of linear structures of spiculated lesions [2]. In their work, the probability density function of the observation vectors for each class is assumed to be normal. we have experienced that the “normal” distribution for each class is nor true. Li *et al.* proposed using a Markov random field model to extract suspicious masses for mass detection [11]. In their study, most of model parameters were chosen empirically, and the mammogram was segmented into three regions (background, fat, and parenchymal or tumors).

Stochastic model-based image segmentation is a technique for partitioning an image into distinctive meaningful regions based on the statistical properties of both gray level and context images. A good segmentation result would depend on suitable model selection for a specific image modality [16], [17] where model selection refers to the determination of both the number of image regions and the local statistical distributions of each region. Furthermore, a segmentation result would be improved

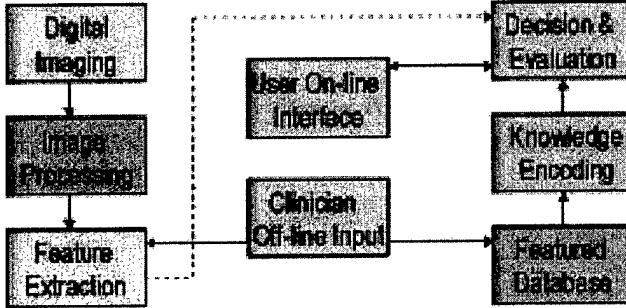


Fig. 1. Major components in CAD.

with preenhanced pattern of interest being segmented. The only assumption for suspected mass site selection is that suspected mass areas should be brighter than the surrounding breast tissues which is valid for most of the real cases. When some masses lie either within an inhomogeneous pattern of fibroglandular tissue or are partially or completely surrounded by fibroglandular tissue, enhancement of mass-related signals is important.

Fig. 1 shows a general block diagram of CAD systems. This paper focuses on “image processing” block, to just automatically pick up all possible lesion sites. We aim on two essential issues in the stochastic model-based image segmentation: enhancement and model selection. Based on the differential geometric characteristics of masses against the background tissues, we propose one type of morphological operation to enhance the mass patterns on mammograms. Then we employ a finite generalized Gaussian mixture (FGGM) distribution to model the histogram of the mammograms where the statistical properties of the pixel images are largely unknown and are to be incorporated. We incorporate the EM algorithm with two information theoretic criteria to determine the optimal number of image regions and the kernel shape in the FGGM model. Finally, we apply a contextual Bayesian relaxation labeling (CBRL) technique to perform the selection of suspected masses. The major differences of our work from the previous work [1]–[6], [8]–[13] are as follows.

- 1) We present a new algorithm of morphological filtering for image enhancement in which the combined operations are applied to the original gray tone image and the higher sensitive lesion site selection of the enhanced images are observed.
- 2) We justify and pilot test the FGGM distribution in modeling mammographic pixel images together with a model selection procedure based on the two information theoretic criteria. This allows an automatic identification of both the number (K) and kernel shape (α) of the distributions of tissue types.
- 3) We develop a new algorithm (CBRL) for segmenting mass areas where the comparable results are achieved as those using Markov random field model-based approaches while with much less computational complexity.

The presentation of this paper is organized as follows. In Section II, the proposed dual morphological operation enhancement technique is described in detail. The theory and algorithm on

FGGM modeling, model selection, and parameter estimation are presented in Section III. This is followed by a discussion on the selection of suspicious masses using the CBRL approach. Evaluation results are given and discussed in Section IV. Finally, the paper is concluded by Section V.

II. MORPHOLOGICAL ENHANCEMENT

One of the main difficulties in suspicious mass segmentation is that mammographic masses are often overlapped with dense breast tissues. Therefore, it is necessary to remove bright background caused by dense breast tissues while preserving the features and patterns related to the masses. For this purpose, background correction is an important step for mass segmentation. We propose a mass pattern-dependent background removal approach using morphological operations.

A. Morphological Filtering Theory

Morphological operations can be employed for many image processing purposes, including edge detection, region segmentation, and image enhancement. The beauty and simplicity of mathematical morphology approach come from the fact that a large class of filters can be represented as the combination of two simple operations: erosion and dilation. Let Z denote the set of integers and $f(i, j)$ denote a discrete image signal, where the domain set is given by $\{i, j\} \in N_1 \times N_2$, $N_1 \times N_2 \subset Z^2$ and the range set by $\{f\} \in N_3$, $N_3 \subset Z$. A structuring element B is a subset in Z^2 with a simple geometrical shape and size. Denote $B^s = \{-b : b \in B\}$ as the symmetric set of B and B_{t_1, t_2} as the translation of B by (t_1, t_2) , where $(t_1, t_2) \in Z^2$. The erosion $f \ominus B^s$ and dilation $f \oplus B^s$ can be expressed as [19]

$$(f \ominus B^s)(i, j) = \min_{t_1, t_2 \in B_{i,j}} (f(t_1, t_2)) \quad (1)$$

$$(f \oplus B^s)(i, j) = \max_{t_1, t_2 \in B_{i,j}} (f(t_1, t_2)). \quad (2)$$

On the other hand, opening $f \circ B$ and closing $f \bullet B$ are defined as [19]

$$(f \circ B)(i, j) = ((f \ominus B^s) \oplus B)(i, j) \quad (3)$$

$$(f \bullet B)(i, j) = ((f \oplus B^s) \ominus B)(i, j). \quad (4)$$

A gray value image can be viewed as a two-dimensional surface in a three-dimensional space. Given an image, the opening operation removes the objects, which have size smaller than the structuring element, with positive intensity. Thus, with the specified structuring element, one can extract different image contexts by taking the difference between the original and opening processed image, which is known as “tophat” operation [19].

B. Morphological Enhancement Algorithms

Based on the properties of morphological filters, we designed one type of mass pattern-dependent enhancement approaches. The algorithm is implemented by dual morphological tophat operations following by a subtraction which is described as follows.

Step 1) The textures without the pattern information of interest are extracted by a tophat operation

$$r_1(i, j) = \max(0, [f(i, j) - (f \circ B_1)(i, j)]) \quad (5)$$

where $f(i, j)$ is the original image, and $r_1(i, j)$ is the residue image between the original image and the opening of the original image by a specified structuring element B_1 . The size of B_1 should be chosen smaller than the size of masses.

Step 2) Let $r_2(i, j)$ be the mass pattern enhanced image by background correction, i.e., by the second tophat operation on $f(i, j)$

$$r_2(i, j) = \max(0, [f(i, j) - (f \circ B_2)(i, j)]) \quad (6)$$

where B_2 is a specified structuring element which has a larger size than masses.

Step 3) The enhanced image $f_1(i, j)$ can be derived as

$$f_1(i, j) = \max(0, [r_2(i, j) - r_1(i, j)]). \quad (7)$$

This operation is called “dual morphological operation.” It can remove the background noise and the structure noise inside the suspected mass patterns. Fig. 2 shows the mass patch and the enhanced results of each step using the dual morphological operation. As we can see from Fig. 2, both background correction [Fig. 2(c)] and dual morphological operation [Fig. 2(d)] enhanced the mass pattern, but dual morphological operation removed more structural noise inside the mass region which in turn would improve the mass segmentation results.

III. MODEL-BASED SEGMENTATION

A. Statistical Modeling

Given a digital image consisting of $N_1 \times N_2$ pixels, assume this image contains K regions. By randomly reordering all pixels in the underlying probability space, one can treat pixel labels as random variables and introduce a prior probability measure π_k . Then the FGGM probability density function (pdf) of gray level of each pixel is given by [17]

$$p(x_i) = \sum_{k=1}^K \pi_k p_k(x_i), \quad i = 1, \dots, N_1 N_2, \\ x_i = 0, 1, \dots, L - 1 \quad (8)$$

where x_i is the gray level of pixel i , and L is the number of gray levels. $p_k(x_i)$ s are conditional region pdfs with the weighting factor π_k , satisfying $\pi_k > 0$, and $\sum_{k=1}^K \pi_k = 1$. The generalized Gaussian pdf given region k is defined by

$$p_k(x_i) = \frac{\alpha \beta_k}{2\Gamma(1/\alpha)} \exp[-|\beta_k(x_i - \mu_k)|^\alpha], \quad \alpha > 0, \\ \beta_k = \frac{1}{\sigma_k} \left[\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)} \right]^{1/2}. \quad (9)$$

where μ_k is the mean, $\Gamma(\cdot)$ is the Gamma function. β_k is a parameter related to the variance σ_k . It can be shown that when

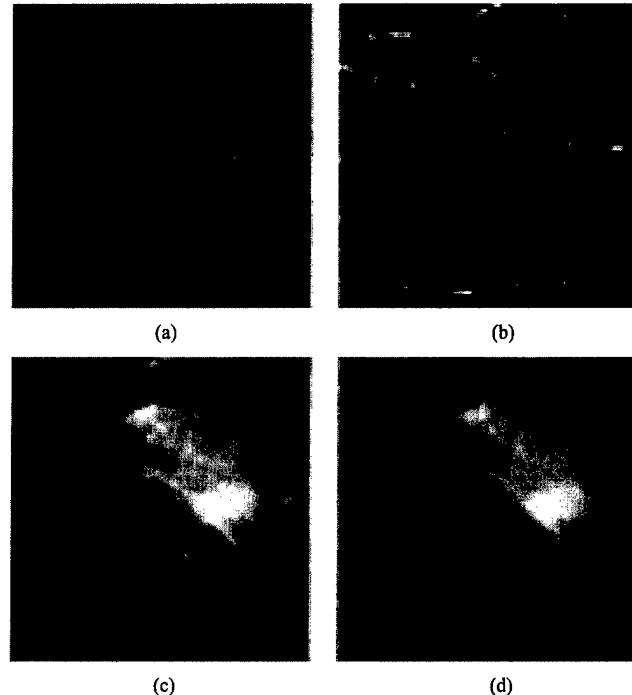


Fig. 2. Original and enhancement result of the mass patch using dual-morphological operation. (a) Original image block $f(i, j)$. (b) Textures $r_1(i, j)$. (c) Background correction result $r_2(i, j)$. (d) Enhanced result $f_1(i, j)$.

$\alpha = 2.0$, one has the Gaussian pdf; when $\alpha = 1.0$, one has the Laplacian pdf. When $\alpha \gg 1$, the distribution tends to a uniform pdf; when $\alpha < 1$, the pdf becomes sharp. Therefore, the generalized Gaussian model is a suitable model to fit the histogram distribution of those images whose statistical properties are unknown since the kernel shape can be controlled by selecting different α values.

The whole image can be well approximated by an independent and identically distributed random field \mathbf{X} . The corresponding joint pdf is

$$P(\mathbf{x}) = \prod_{i=1}^{N_1 N_2} \sum_{k=1}^K \pi_k p_k(x_i) \quad (10)$$

where $\mathbf{x} = [x_1, x_2, \dots, x_{N_1 N_2}]$, and $\mathbf{x} \in \mathbf{X}$. $p_k(x_i)$ is given in (9). Based on the joint probability measure of pixel images, the likelihood function under FGGM modeling can be expressed as $\mathcal{L}(\mathbf{r}) = \prod_{i=1}^{N_1 N_2} p_{\mathbf{r}}(x_i)$ where $\mathbf{r} : \{K, \alpha, \pi_k, \mu_k, \sigma_k, k = 1, \dots, K\}$ denotes the model parameter set.

B. Model Identification

With an appropriate system likelihood function, the objective of model identification is to estimate the model parameters by maximizing the likelihood function, or equivalently minimizing the relative entropy between the image histogram $p_{\mathbf{x}}(u)$ and the estimated pdf $p_{\mathbf{r}}(u)$, where u is the gray level. Based on the FGGM model, the EM algorithm is applied to estimate the model parameters. The EM algorithm is an iterative technique for maximum-likelihood (ML) estimation [20]. Recently, it has been used in many medical imaging applications [15]. Instead

of evaluating directly the value of ML, we use the global relative entropy (GRE) between the histogram and the estimated FGGM distribution to measure the performance of parameter estimation, given by

$$\text{GRE}(p_{\mathbf{x}} \| p_r) = \sum_u p_{\mathbf{x}}(u) \log \frac{p_{\mathbf{x}}(u)}{p_r(u)}. \quad (11)$$

Motivated by the same spirit of conventional EM algorithm for finite normal mixtures (FNMs), we formulated the EM algorithm to estimate the parameter values of the FGGM. The algorithm is summarized as follows.

EM Algorithm:

- 1) For $\alpha = \alpha_{\min}, \dots, \alpha_{\max}$
 - $m = 0$, given initialized $\mathbf{r}^{(0)}$
 - E-step: for $i = 1, \dots, N_1 N_2$, $k = 1, \dots, K$, compute the probabilistic membership

$$z_{ik}^{(m)} = \frac{\pi_k^{(m)} p_k(x_i)}{\sum_{k=1}^K \pi_k^{(m)} p_k(x_i)}. \quad (12)$$

- M-step: for $k = 1, \dots, K$, compute the updated parameter estimates

$$\left\{ \begin{array}{l} \pi_k^{(m+1)} = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1 N_2} z_{ik}^{(m)} \\ \mu_k^{(m+1)} = \frac{1}{N_1 N_2 \pi_k^{(m+1)}} \sum_{i=1}^{N_1 N_2} z_{ik}^{(m)} x_i \\ \sigma_k^{2(m+1)} = \frac{1}{N_1 N_2 \pi_k^{(m+1)}} \sum_{i=1}^{N_1 N_2} z_{ik}^{(m)} (x_i - \mu_k^{(m+1)})^2. \end{array} \right. \quad (13)$$

- When $|\text{GRE}^{(m)}(p_{\mathbf{x}} \| p_r) - \text{GRE}^{(m+1)}(p_{\mathbf{x}} \| p_r)| \leq \epsilon$ is satisfied, go to Step 2 Otherwise, $m = m + 1$ and go to E-Step.

- 2) Compute GRE, and go to Step 1.
- 3) Choose the optimal $\hat{\mathbf{r}}$ which corresponds to the minimum GRE.

As we mentioned in Section I, the two important parameters in model selection are K and α . Determination of the region parameter K directly affects the quality of the resulting model parameter estimation and in turn, affects the result of segmentation. In this paper we propose an approach to determine the value of K based on two popular information theoretic criteria introduced by Akaike [23] and by Rissanen [24]. Akaike proposed to select the model that gives the minimum Akaike information criterion (AIC), defined by

$$\text{AIC}(K) = -2 \log(\mathcal{L}(\hat{\mathbf{r}}_{ML})) + 2K' \quad (14)$$

where $\hat{\mathbf{r}}_{ML}$ is the ML estimate of the model parameter set \mathbf{r} , and K' is the number of free adjustable parameters in the model [15], [23]. AIC criterion will select the correct number of the image regions K_0 when

$$K_0 = \arg \left\{ \min_{1 \leq K \leq K_{\max}} \text{AIC}(K) \right\}. \quad (15)$$

Rissanen addressed the problem from a quite different point of view. Rissanen reformulated the problem explicitly as an information coding problem in which the best model fitness is measured such that it assigns high probabilities to the observed data while at the same time the model itself is not too complex to describe [24]. The model is selected by minimizing the total description length defined by minimum description length (MDL)

$$\text{MDL}(K) = -\log(\mathcal{L}(\hat{\mathbf{r}}_{ML})) + 0.5K' \log(N_1 N_2). \quad (16)$$

Similarly, the correct number of the distinctive image regions K_0 will be estimated when

$$K_0 = \arg \left\{ \min_{1 \leq K \leq K_{\max}} \text{MDL}(K) \right\}. \quad (17)$$

C. Bayesian Relaxation Labeling

Once the FGGM model is given, a segmentation problem is the assignment of labels to each pixel in the image. A straightforward way is to label pixels into different regions by maximizing the individual likelihood function $p_k(x)$. This approach is called ML classifier, which is equivalent to a multiple thresholding method. Usually, this method may not achieve a good performance since there is lack of local neighborhood information to be included to make a good decision. CBRL algorithm [25] is one of the approaches, which can incorporate the local neighborhood information into labeling procedure and thus improve the segmentation performance. In this study, we developed the CBRL algorithm to perform/refine pixel labeling based on the localized FGGM model, which is defined as follows.

Let ∂i be the neighborhood of pixel i with an $m \times m$ template centered at pixel i . An indicator function is used to represent the local neighborhood constraints $R_{ij}(l_i, l_j) = I(l_i, l_j)$, where l_i and l_j are labels of pixels i and j , respectively. Note that pairs of labels are now either compatible or incompatible. Similar to reference [25], one can compute the frequency of neighbors of pixel i which has the same label values k as at pixel i

$$\pi_k^{(i)} = p(l_i = k | \mathbf{l}_{\partial i}) = \frac{1}{m^2 - 1} \sum_{j \in \partial i, j \neq i} I(k, l_j) \quad (18)$$

where $\mathbf{l}_{\partial i}$ denotes the labels of the neighbors of pixel i . Since $\pi_k^{(i)}$ is a conditional probability of a region, the localized FGGM pdf of gray level x_i at pixel i is given by

$$p(x_i | \mathbf{l}_{\partial i}) = \sum_{k=1}^K \pi_k^{(i)} p_k(x_i) \quad (19)$$

where $p_k(x_i)$ is given in (9). Assuming gray values of the image are conditional independent, the joint pdf of \mathbf{x} , given the context labels \mathbf{l} , is

$$P(\mathbf{x} | \mathbf{l}) = \prod_{i=1}^{N_1 N_2} \sum_{k=1}^K \pi_k^{(i)} p_k(x_i) \quad (20)$$

where $\mathbf{l} = (l_i : i = 1, \dots, N_1 N_2)$.

It is known that CBRL algorithm can obtain a consistent labeling solution based on the localized FGGM model (19). Since

TABLE I
DISTRIBUTION OF THE EFFECTIVE SIZE OF THE 186 MASSES USED IN THIS STUDY. THE EFFECTIVE SIZE IS DEFINED AS THE SQUARE ROOT OF THE PRODUCT OF THE MAXIMUM AND MINIMUM DIAMETERS OF THE MASS

	0 – 5mm	6 – 10mm	11 – 15mm	16 – 20mm	21 – 25mm	26 – 30mm
#	3	55	78	29	17	4

l represents the labeled image, it is consistent if $S_i(l_i) \geq S_i(k)$, for all $k = 1, \dots, K$ and for $i = 1, \dots, N_1N_2$ [25], where

$$S_i(k) = \pi_k^{(i)} p_k(x_i). \quad (21)$$

Now we can define

$$A(l) = \sum_{i=1}^{N_1N_2} \left(\sum_k I(l_i, k) S_i(k) \right) \quad (22)$$

as the average measure of local consistency, and

$$LC_i = \sum_k I(l_i, k) S_i(k), \quad i = 1, \dots, N_1N_2 \quad (23)$$

represents the local consistency based on l . The goal is to find a consistent labeling l which can maximize (22). In the real application, each local consistency measure LC_i can be maximized independently. In [25], it has been shown that when $R_{ij}(l_i, l_j) = R_{ji}(l_j, l_i)$, if $A(l)$ attains a local maximum at l , then l is a consistent labeling.

Based on the localized FGGM model, $l_i^{(0)}$ can be initialized by ML classifier

$$l_i^{(0)} = \arg \left\{ \max_k p_k(x_i) \right\}, \quad k = 1, \dots, K. \quad (24)$$

Then, the order of pixels is randomly permuted and each label l_i is updated to maximize LC_i , i.e., classify pixel i into k th region if

$$l_i = \arg \left\{ \max_k \pi_k^{(i)} p_k(x_i) \right\}, \quad k = 1, \dots, K \quad (25)$$

where $p_k(x_i)$ is given in (9), $\pi_k^{(i)}$ is given in (18). By considering (24) and (25), we developed a modified CBRL algorithm as follows.

CBRL Algorithm:

- 1) Given $l^{(0)}$, $m = 0$
- 2) Update pixel labels
 - Randomly visit each pixel for $i = 1, \dots, N_1N_2$
 - Update its label l_i according to

$$l_i^{(m)} = \arg \left\{ \max_k \pi_k^{(i)(m)} p_k(x_i) \right\}.$$

- 3) When

$$\frac{\sum (l^{(m+1)} \oplus l^{(m)})}{N_1N_2} \leq 1\%,$$

stop; otherwise, $m = m + 1$, and repeat Step 2.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the results of using the morphological filtering and model-based segmentation approach we have introduced for enhancement and segmentation of suspi-

cious masses in mammographic images. In addition to the qualitative assessment by the radiologists, we introduce several objective measures to assess the performance of the algorithms we have proposed for enhancement and segmentation.

A testing data set of 200 mammograms and two simulated tone images were used to test and evaluate the performance of the algorithms in this study. The mammograms were selected from the Mammographic Image Analysis Society (MIAS) database and the Brook Army Medical Center (BAMC) database created by the Department of Radiology at Georgetown University Medical Center. Of the 200 mammograms, 50 mammograms are normal, and each of the 150 abnormal mammograms contains at least one mass case of varying size, subtlety, and location. The areas of suspicious masses were identified by an expert radiologist based on visual criteria and biopsy proven results. The total data set includes 113 benign and 73 malignant masses. The distribution of the masses in terms of size is shown in Table I. The BAMC films were digitized with a laser film digitizer (Lumiscan 150) at a pixel size of $100 \mu\text{m} \times 100 \mu\text{m}$ and 4096 gray levels (12 bits). Before the method was applied the digital mammograms were smoothed by averaging 4×4 pixels into one pixel. According to radiologists, the size of small masses is 3–15 mm in effective diameter. A 3-mm object in an original mammogram occupies 30 pixels in a digitized image with a $100\text{-}\mu\text{m}$ resolution. After reducing the image size by four times, the object will occupy the range of about seven to eight pixels. The object with the size of seven pixels is expected to be detectable by any computer algorithm. Therefore, the shrinking step is applicable for mass cases and can save computation time.

Experimental Evaluation of Morphological Enhancement: In order to justify the suitability of morphological structural elements, the geometric properties of the contexts and textures in mammograms were studied. The basic idea is to keep all mass-like objects within certain size range and remove all others by using the proposed morphological filters with specific structural elements. At the resolution of $400 \mu\text{m}$, a disk with a diameter of seven pixels was chosen as the morphological structuring elements B_1 to extract textures in mammograms. Since the smallest masses have seven pixels in diameter with the resolution of $400 \mu\text{m}$, this procedure would not destroy mass information. For the purpose of background correction, a disk with a diameter of 75 pixels was used as the morphological structuring element B_2 . An object with a diameter of 75 pixels corresponds to 30 mm in the original mammogram. This indicates that all masses with sizes up to 30 mm can be enhanced by background correction. Masses larger than 30 mm are rare cases in the clinical setting. In the last stage of our approach, we applied morphological opening and closing filtering using a disk with a diameter of five to eliminate small objects which also contribute to texture noise.

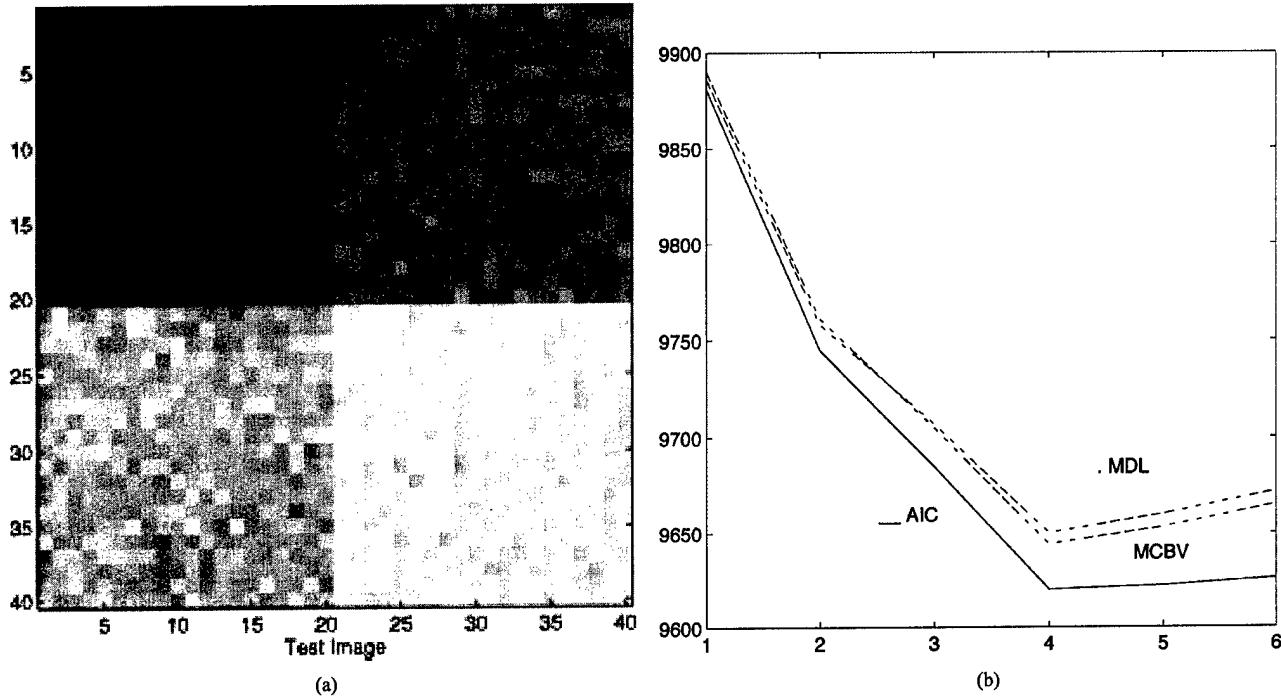


Fig. 3. (a) Original simulated test image for model selection ($k_0 = 4$, SNR = 10 dB) and (b) the AIC/MDL curves in model selection ($\sigma = 30$).

All testing mammograms were processed using the proposed enhancement approach with the suggested structuring element B_1 and B_2 . Fig. 5 shows processed mammogram examples using the morphological enhancement. Compared the enhanced results [Fig. 5(b) and (d)] with the original mammograms [Fig. 5(a) and (c)], the proposed method not only enhanced all suspected mass patterns and reduced the texture noise, but also removed the background noise. In summary, the proposed morphological enhancement approach can enhance mass patterns and remove texture structure noises. For dense mammograms, such as the second example in Fig. 5(c) and (d), the mass is obscured by dense fibroglandular tissues, our experience shows applying the dual morphological operation to remove the fibroglandular tissue background is useful. In addition to the visual evaluation by the radiologist, we performed the segmentation to assess the effectiveness of the morphological filtering, based on the enhanced mammograms and the original mammograms.

Simulated Evaluation of Segmentation Algorithms: The performance of model selection using two frequently used methods, i.e., the AIC and MDL [22], were first tested and compared in the simulation study. The computer-generated data was made up of four overlapping normal components. Each component represents one local region. The value for each component were set to a constant value, the noise of normal distribution was then added to this simulation digital phantom. Three noise levels with different variance were set to keep the same signal-to-noise ratio (SNR), where SNR is defined by

$$\text{SNR} = 10 \log_{10} \frac{(\Delta\mu)^2}{\sigma^2} \quad (26)$$

where $\Delta\mu$ is the mean difference between regions, and σ^2 is the noise power. The original data for the simulation study are

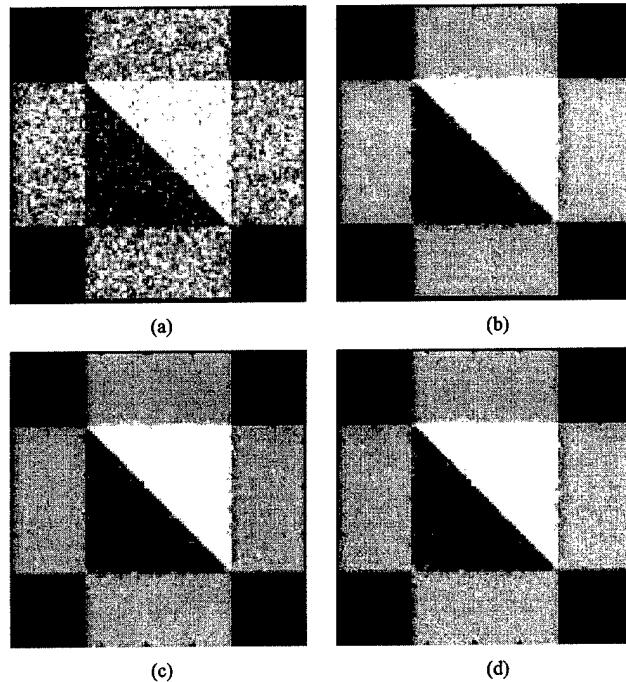


Fig. 4. Image segmentation by CBRL on simulated image (with initialization by ML classification). (a) ML initialization. (b) First iteration in CBRL. (c) Second iteration in CBRL. (d) Third iteration in CBRL.

TABLE II
COMPARISON OF CBRL, ICM, AND MICM ALGORITHM: SIMULATED DATA

Item	CBRL Result	ICM Result	MICM Result
Classification Error	0.7935%	0.7508%	0.3113%

given in Fig. 3(a). The AIC and MDL curves, as functions of the number of local clusters K , are plotted in Fig. 3(b). According

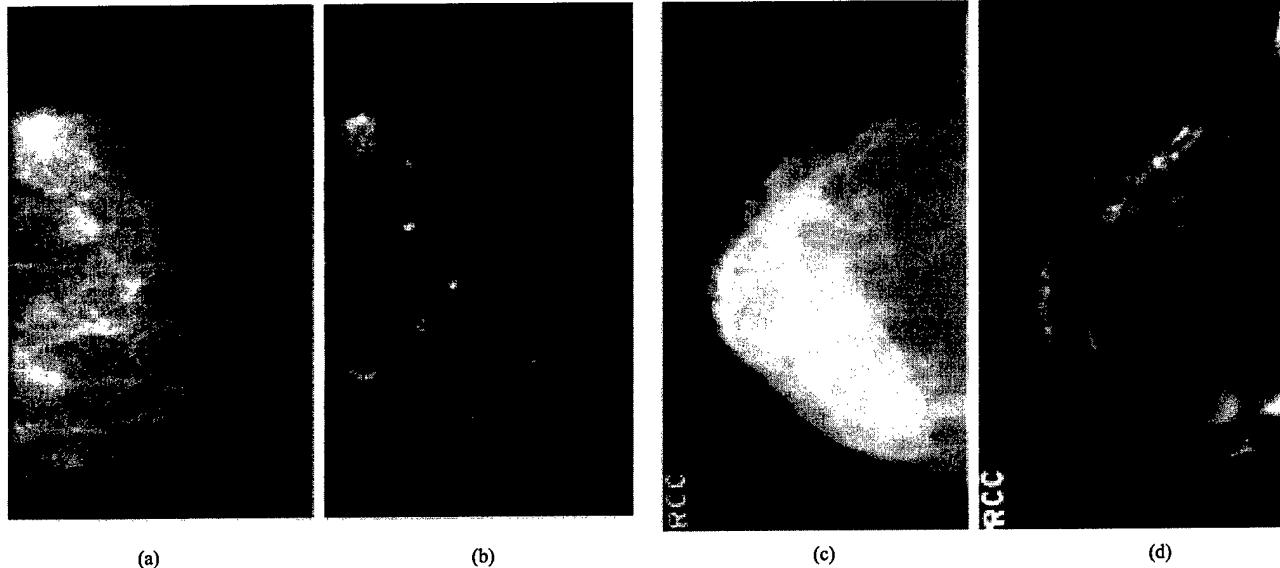


Fig. 5. Examples of mass enhancement. (a) Original mammogram. (b) Enhanced mammogram. (c) and (d) Another original mammogram and its enhanced result.

to the information theoretic criteria, the minima of these curves indicate the correct number of the local regions. From this experimental figure, it is clear that the number of local regions suggested by these criteria are all correct.

For the validation of image segmentation using CBRL, we apply the algorithm first to a simulated image. We use ML classifier to initialize image segmentation, i.e., to initialize the quantified image by selecting the pixel label with largest likelihood at each node. The classification error after initialization is uniformly distributed over the spatial domain as shown in Fig. 4(a). Our experience suggested this to be a very suitable starting point for contextual relaxation labeling [21]. The CBRL is then performed to fine tune the image segmentation. It should be emphasized that the ground truth is known in this simulated experiment, the percentage of total classification error is used as the criterion for evaluating the performance of segmentation technique. In Fig. 4(a)–(d), the initial segmentation by the ML classification and the stepwise results of three iterations in the CBRL are presented. In this experiment, algorithm initialization results in an average classification error of 30%. It can be clearly seen that a dramatic improvement is obtained after several iterations of the CBRL by using local constraints determined by the context information. In addition, the convergence is fast as one can see, after the first iteration most of the misclassification are removed. We have also implemented two other independent and popular algorithms, namely, the iterated conditional mode (ICM) and the modified iterated conditional mode (MICM) algorithms, so as to assess the comparative performance of the segmentation results among different approaches [21], [22]. The only assumption being made by these three methods is the Markovian property of the context images which can be well justified by the underlying cell oncology and pathology. We have applied these three algorithms to the same testing image and the corresponding classification errors are presented in Table II. The final percentage of classification errors for Fig. 4(d) is 0.7935%. From this experimental comparison, it can be concluded that three algorithms achieved com-

TABLE III
COMPUTED AICS FOR THE FGGM MODEL WITH DIFFERENT α

K	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$
2	651250	650570	650600	650630
3	646220	644770	645280	646200
4	645760	644720	645260	646060
5	645760	644700	645120	646040
6	645740	644670	645110	645990
7	645640	644600	645090	645900
8	645550(min)	644570(min)	645030(min)	645850(min)
9	645580	644590	645080	645880
10	645620	644600	645100	645910

TABLE IV
COMPUTED MDLS FOR THE FGGM MODEL WITH DIFFERENT α

K	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$
2	651270	650590	650630	650660
3	646260	644810	645360	646350
4	645860	644770	645280	646150
5	645850	644770	645280	646100
6	645790	644750	645150	646090
7	645720	644700	645120	645930
8	645680(min)	644690(min)	645100(min)	645900(min)
9	645710	644710	645140	645930
10	645790	644750	645180	645960

parable segmentation accuracy and the result produced by the MICM algorithm is most superior, though in terms of computational complexity the CBRL algorithm is the least. It should be noticed that since in MICM algorithm an inhomogeneous configuration of the Markov random field is used, its superior performance is reasonable.

On Model-Based Segmentation—Real Case Study: In the real case study, we used two information criteria (AIC and MDL) to determine K . Tables III and IV shows the AIC and MDL values with different K and α of the FGGM model based on one original mammogram. As it can be seen from Tables III and IV, although with different α , all AIC and MDL values achieve the minimum when $K = 8$. It indicates that AIC and

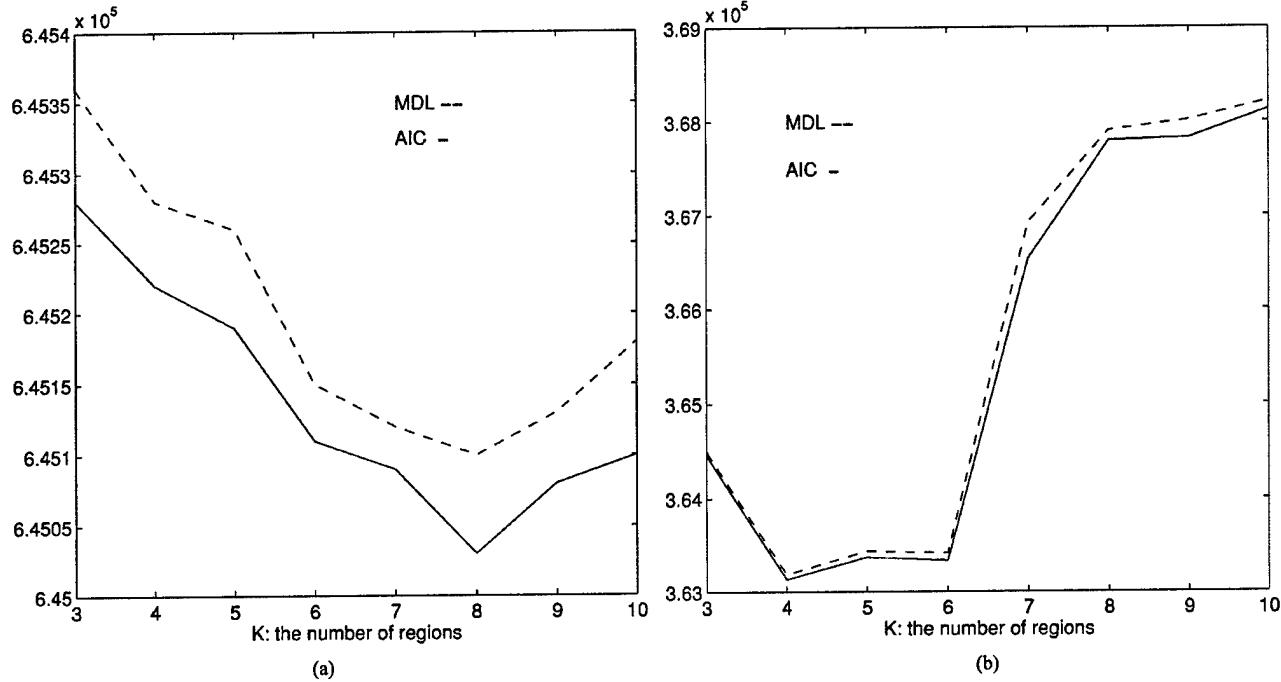


Fig. 6. AIC and MDL curves with different number of region K . (a) Result based on the original mammogram, the optimal $K = 8$. (b) Result based on the enhanced mammogram, the optimal $K = 4$.

MDL are relatively insensitive to the change of α . With this observation, we can decouple the relation between K and α and choose the appropriate value of one while fixing the value of another. Fig. 6(a) and (b) are two examples of AIC and MDL curves with different K and fixed $\alpha = 3.0$. Fig. 6(a) is based on the original mammogram and Fig. 6(b) is based on the enhanced mammogram. As we can see in Fig. 6(a), both criteria achieved the minimum when $K = 8$. It should be noticed that though no ground truth is available in this case, our extensive numerical experiments have shown a very consistent performance of the model selection procedure and all the conclusions were strongly supported by the previous independent work reported by [14]. Fig. 6(b) indicates that $K = 4$ is the appropriate choice for the mammogram enhanced by dual morphological operation. This is believed to be reasonable since the number of regions decrease after background correction.

We fixed $K = 8$, and changed the value of α for estimating the FGGM model parameters using the proposed EM algorithm with the original mammogram. The GRE value between the histogram and the estimated FGGM distribution was used as a measure of the estimation bias. We found that GRE achieved a minimum distance when the FGGM parameter $\alpha = 3.0$ as shown in Fig. 7. The similar result was shown when we applied the EM algorithm to the enhanced mammogram with $K = 4$. This indicated that the FGGM model might be better than the FNM model ($\alpha = 2.0$) in modeling mammographic images when the true statistical properties of mammograms are generally unknown, though the FNM has been most often chosen in many previous work [15].

After the determination of all model parameters, every pixel of the image was labeled to a different region (from 1 to K) based on the CBRL algorithm. We then selected the brightest re-

TABLE V
COMPARISON OF SEGMENTATION ERROR RESULTING FROM NONCONTEXTUAL AND CONTEXTUAL METHODS

Method	Soft Classification	Bayesian Classification	CBRL
GRE Value	0.0067	0.4406	0.1578

gion, which corresponding to label K , plus a criterion of closed isolated area, as the candidate region of suspicious masses. According to the visual inspections by the radiologists, when we use $K - 1$ instead of K , the results are over-segmented. For the case of using $K + 1$, the results are under-segmented. In order to quantify the performance differences between the different segmentation methods, several groups have suggested that the segmentation results may be compared against radiologists' outlines of the lesions [3]. Though the proposed comparison measures are quantitative, the performance measures are still qualitative, since the reference base (e.g., gold standard by the radiologists) is qualitative, subjective, and imperfect. Therefore, in this model-supported approach, in addition to the visual inspections by the radiologists, we have also introduced an objective measure, the GRE between the histogram of the pixel images $p_x(u)$ and the FGGM of the segmented image $p_{x,l}(u)$ to assess the performance of the segmentation, defined by

$$\text{GRE}(p_x(u) \| p_{x,l}(u)) = \sum_u p_x(u) \log \frac{p_x(u)}{p_{x,l}(u)} \quad (27)$$

where l is the context image estimated by the segmentation algorithm. Considering that the ergodic theorem is the most fundamental principle in the detection and estimation theory, it is believed that when a good segmentation is achieved, the distance between the $p_x(u)$ and $p_{x,l}(u)$ should be minimized and

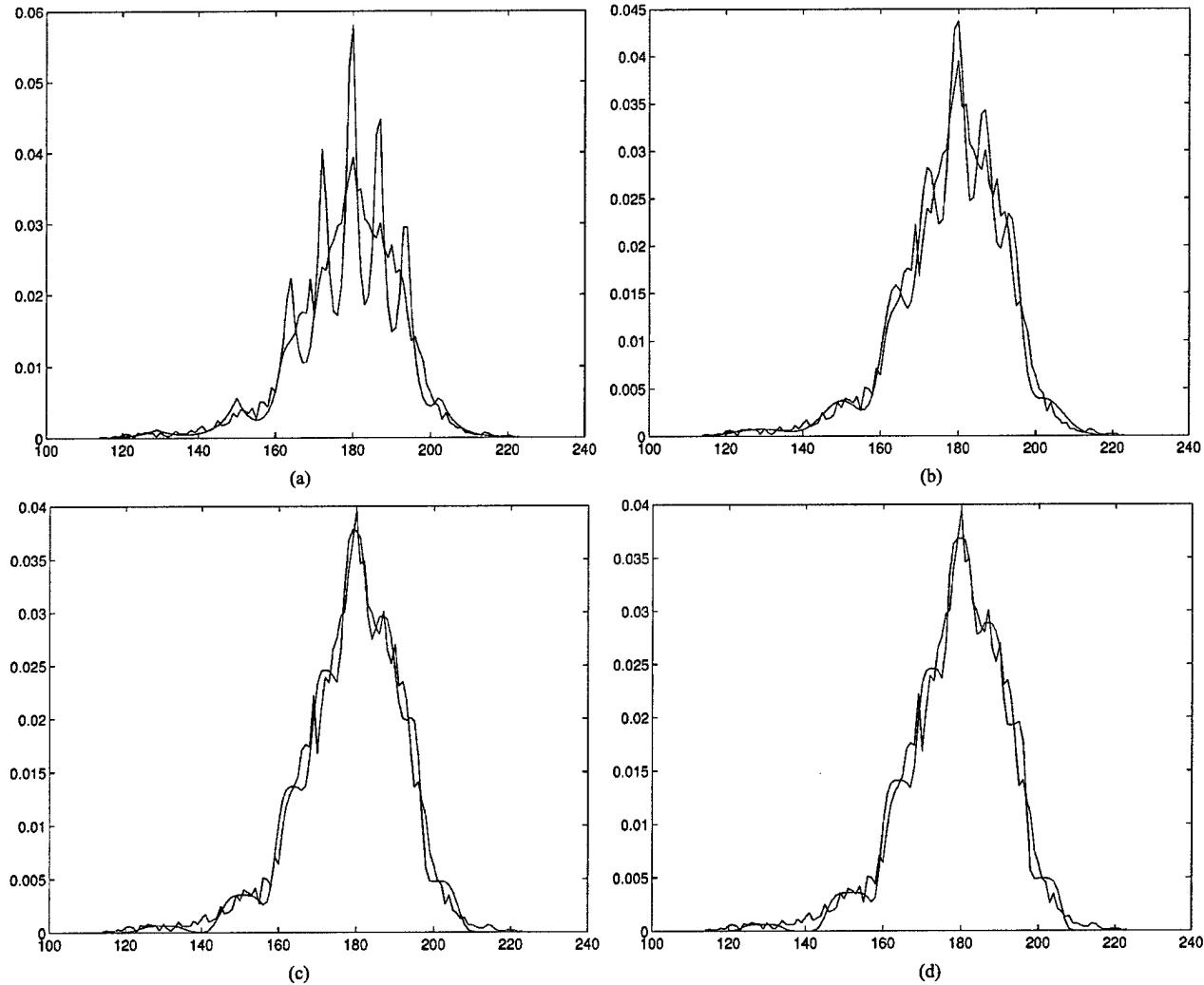


Fig. 7. Comparison of learning curves and histogram of the original mammogram with different α , $k = 8$. The optimal $\alpha = 3.0$. (a) $\alpha = 1.0$, GRE = 0.0783. (b) (a) $\alpha = 2.0$, GRE = 0.0369. (c) $\alpha = 3.0$, GRE = 0.0251. (d) $\alpha = 4.0$, GRE = 0.0282.

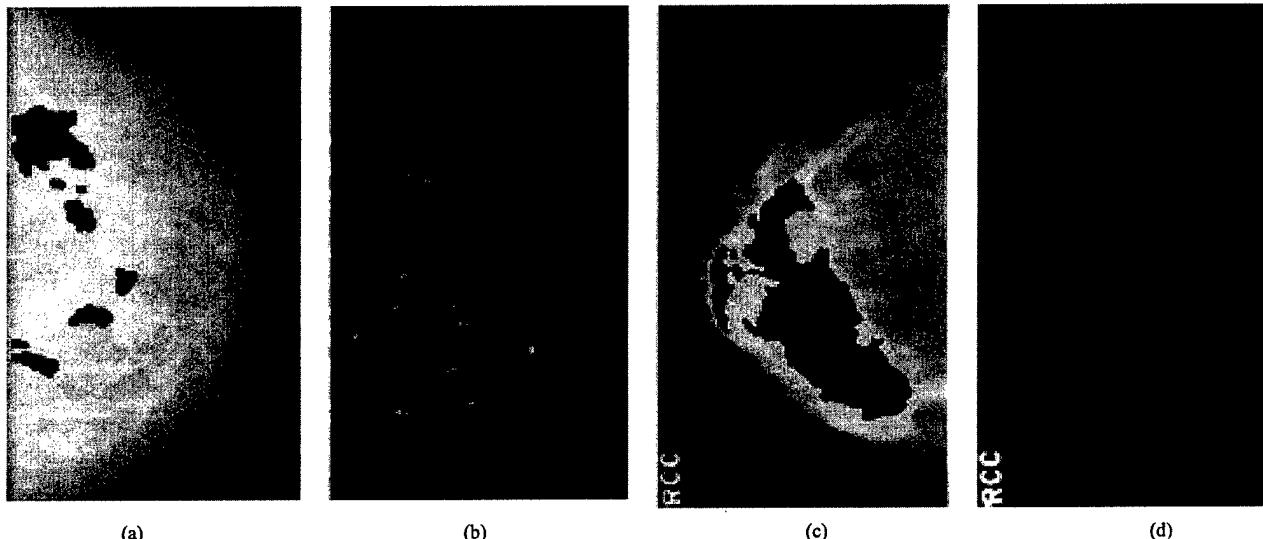


Fig. 8. Suspected mass segmentation results based on the original mammogram. (b) Result based on the enhanced mammogram, $K = 4$, $\alpha = 3.0$. (c) and (d) Results based on another original mammogram and its enhanced image.

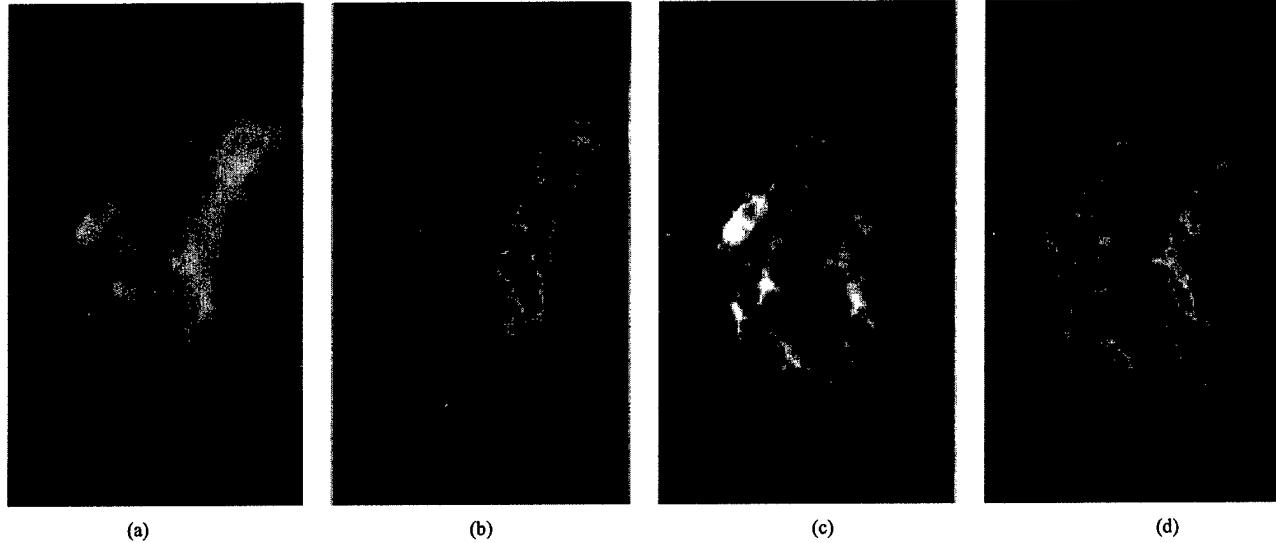


Fig. 9. Examples of normal mixed fatty and glandular mammogram. (a) Original mammogram. (b) Segmentation result based on the original mammogram. (c) Enhanced mammogram. (d) Result based on the enhanced mammogram, $k = 4$, $\alpha = 3.0$.

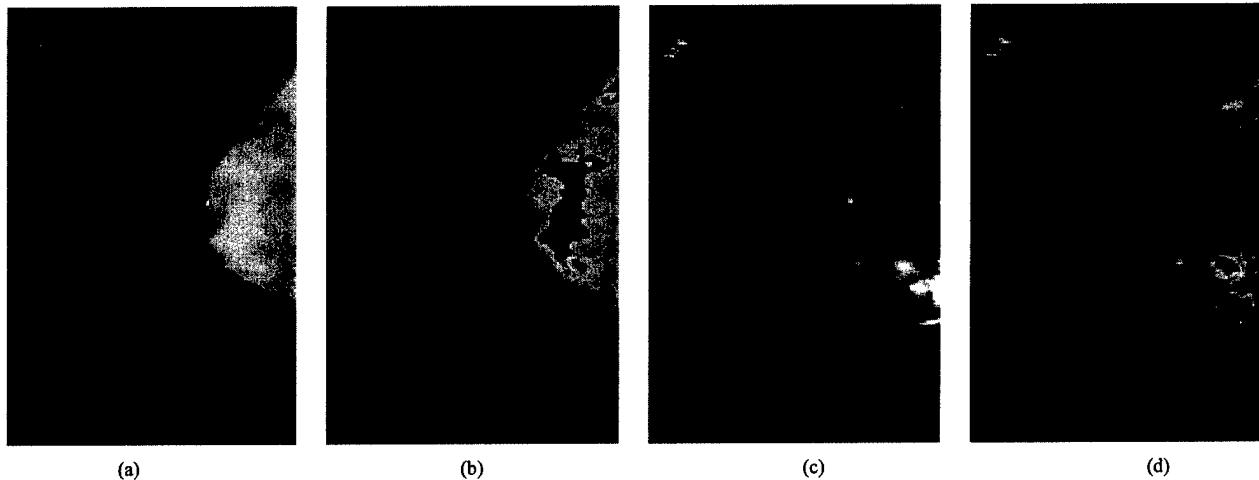


Fig. 10. Examples of normal dense mammogram. (a) Original mammogram. (b) Segmentation result based on the original mammogram. (c) Enhanced mammogram. (d) Result based on the enhanced mammogram, $k = 4$, $\alpha = 3.0$.

this measure links the image text and its sample averages. Our experience has suggested that this post-segmentation measure may be a suitable objective criterion for evaluating the quality of image segmentation in a fully unsupervised situation [22], [26]–[28]. Table V shows our evaluation data from three different segmentation methods when applied to the real images.

Performance of Combined Morphological Filtering and Model-Based Segmentation using a Larger Database: The proposed segmentation method was used to extract suspicious mass regions from the 200 testing mammograms. Without enhancement, a total of 1142 potential mass regions were isolated including 114 of the 186 true masses. With enhancement, a total of 3143 potential mass regions were extracted including 181 of the 186 true masses. The results demonstrated that more true masses were picked up after enhancement although more false cases were also included. The undetected areas mainly occurred at the lower intensity side of the shaded objects or obscured by fibroglandular tissues that, however, were extracted on morpho-

logical enhanced mammograms. In addition, when the margins of masses are ill defined, only parts of suspicious masses were extracted from the original mammograms. For the purpose of “lesion site selection,” we believe that the sensitivity should be the sole criterion for the performance evaluation of the method. We have 181/186 versus 114/186. Our method is unsupervised and automatic and does not involve any detection effort at this moment. To our best knowledge, there is no objective criterion available for the evaluation of image enhancement performance before a detection effort is involved. We only claimed that the enhancement step is important and effective with respect to the purpose of “lesion site selection.”

Fig. 8 demonstrates some segmentation results based on the original and enhanced mammograms. We compared the segmentation results based on the enhanced mammogram ($K = 4$, and $\alpha = 3.0$) with those based on the original mammogram ($K = 8$, and $\alpha = 3.0$) as shown in Fig. 8. Comparing the results in Fig. 8(b) with those in Fig. 8(a), we can see that after

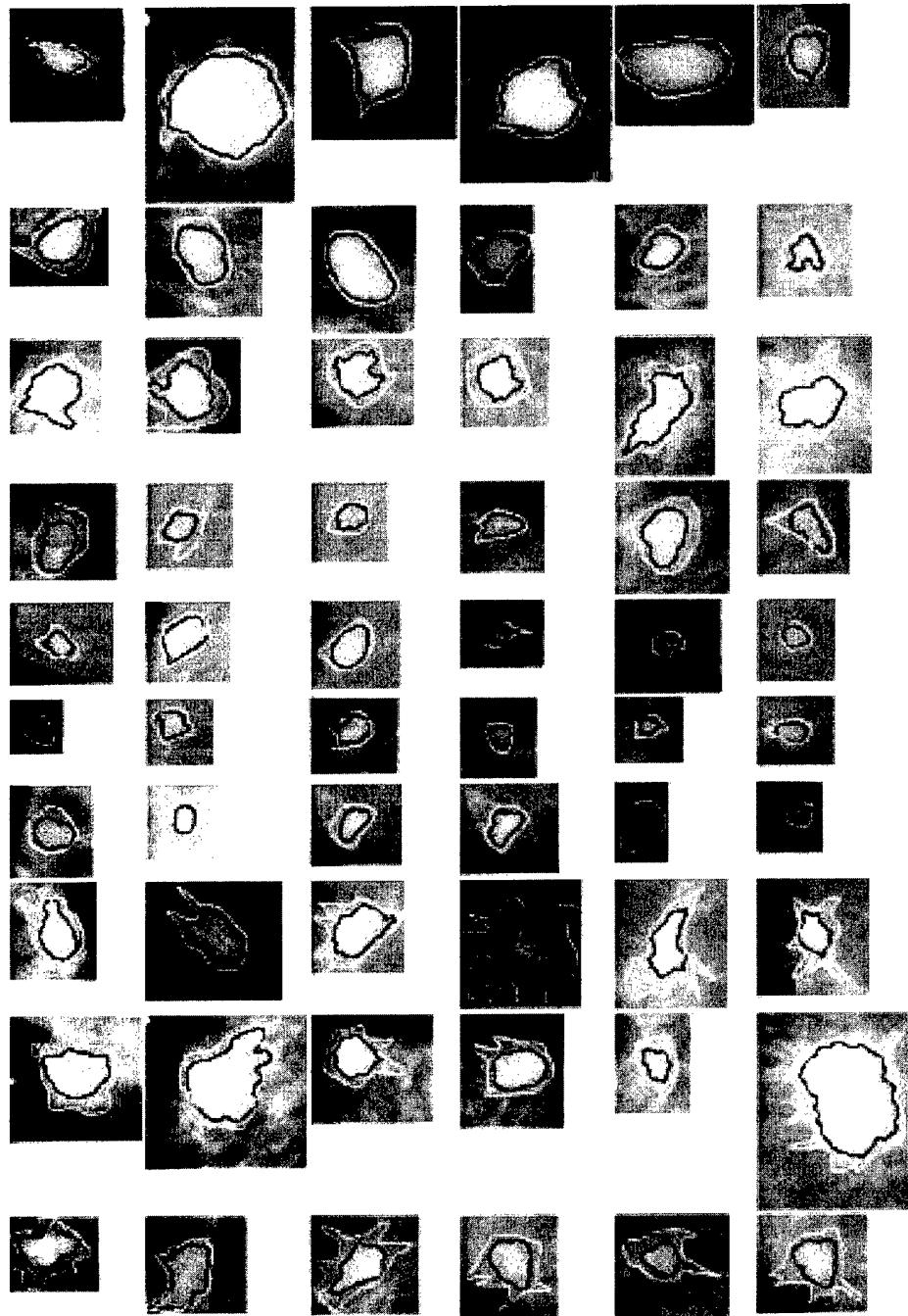


Fig. 11. Comparison results of segmentation based on the enhanced mammograms. Black outlines denote the computer-segmented result. White outlines denote the radiologist-segmented results.

enhancement, a more accurate region was detected for the suspected mass which has ill-defined margin. Getting an accurate suspected region is a crucial issue since geometric features are extracted based on suspected regions and these features are very important for further true mass detection. In addition, we observed that one suspected mass was missed in Fig. 8(a) but was detected in Fig. 8(b). As we have mentioned in Section I, none of the suspected masses should be missed in the segmentation step. Fig. 8(c) and (d) demonstrate the segmentation of a suspected

mass that lies in dense breast tissue. As shown in Fig. 8(c), the whole fibroglandular tissue area was segmented when based on the original mammogram. After enhancement, the suspected region was segmented exactly as shown in Fig. 8(d).

We have also included the segmentation results on the normal mammograms. Fig. 9 demonstrate the segmentation results based on the original and enhanced mixed fatty and glandular mammograms. Fig. 10 demonstrate the segmentation results based on the original and enhanced dense mammograms. We

would like to emphasize that the objective of this paper is to provide a segmentation technique which can enhance and extract potential mass site from the background so that the characterization of the related mass pattern can be accurately extracted in terms of focused feature selection and analysis. The method of course will produce many mass-like areas, but it will be a plausible outcome since the accurate description of nonmass cases characterized by mass-like sites will benefit the follow-on detection step where the performance of the classifier depends on an accurate separation of mass and nonmass in the featured spaces. The details will be described in [29].

For the purpose of evaluating the performance of the segmentation method, we used both simulated studies and expert visual inspection to validate the methods and results. The radiologist has concluded that the lesion characteristics after the proposed enhancement have been better displayed and all possible lesion areas have been successfully identified. In addition to the visual inspection, we have measured the overlap between the computer-segmented and the radiologist segmented mass regions to evaluate our method. Fig. 11 shows the comparison results of segmentation based on the enhanced mammograms. Fig. 11 includes 60 benign and malignant mass patches which were cut from the whole mammograms after the segmentation. The white outline was drawn by the radiologist while the black outline was produced by the computer and was superimposed upon the original image. As we can see from Fig. 11, for most of cases, the ratio of mutual overlap area of the radiologist segmented mass region and the computer-segmented mass region to the radiologist segmented mass area is large than 50%. In addition, even the poorest result picked the true lesion in the correct location and depicted the characteristics of the mass reasonably. It is important to understand that "lesion area segmentation" is not our objective, so there is no "best" or "worst" segmentation results. Our objective is "lesion site selection" with a possible highest sensitivity through a global unsupervised enhancement and segmentation scheme.

V. CONCLUSION

In this paper, we propose a combined method of using morphological operations, a FGGM modeling, and a CBRL to enhance and segment various breast tissue textures and suspicious mass lesions from mammographic images. This phase is a crucial step in mass detection for an improved CAD. We emphasized the importance of model selection which includes the selection of the number of image regions K and the selection of FGGM kernel shape controlled by α . The experimental results indicate that the suspected mass sites selection can be affected by different K and α . We proposed the EM algorithm together with the information theoretic criteria to determine the optimal K and α . With optimal K and α , the segmentation results can be significantly improved. We also showed that with the proposed pattern-dependent enhancement algorithm using morphological operations, the subtle masses can be segmented more accurately than those when the original image is used for extraction without enhancement. To summarize, the morphological filtering enhancement combined with the stochastic model-based segmentation is an effective way to extract mammographic suspicious

patterns of interest, and thereby may facilitate the overall performance of mammographic CAD of breast cancer.

ACKNOWLEDGMENT

The authors would like to thank Z. Gu of the Lombardi Cancer Center and I. Sesterhenn of the Armed Forces Institute of Pathology for their scientific input on the knowledge of cell oncology and pathology, and R. Shah MD, Director of Breast Imaging, BAMC for his evaluation of cases to our database.

REFERENCES

- [1] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized detection of malignant tumors on digital mammograms," *IEEE Trans. Med. Imag.*, vol. 18, pp. 369–378, May 1999.
- [2] R. Zwigelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. M. Boggis, "Model-based detection of spiculated lesions in mammograms," *Med. Image Anal.*, vol. 3, no. 1, pp. 39–62, 1999.
- [3] M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Trans. Med. Imag.*, vol. 17, pp. 510–517, Aug. 1998.
- [4] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammogram," *IEEE Trans. Med. Imag.*, vol. 15, pp. 611–619, Oct 1996.
- [5] W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms," *Excerpta Medica*, vol. 1119, pp. 411–415, 1996.
- [6] N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Trans. Med. Imag.*, vol. 15, no. 1, pp. 59–67, 1996.
- [7] M. Sameti and R. K. Ward, "A fuzzy segmentation algorithm for mammogram partition," in *Digital Mammography*, ser. International Congress Series, K. Doi, Ed. Amsterdam, The Netherlands: Elsevier, 1996, pp. 471–474.
- [8] W. P. Kegelmeyer Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," *Radiology*, vol. 191, pp. 331–337, 1994.
- [9] F. F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Investigat. Radiol.*, vol. 28, no. 6, pp. 473–481, 1993.
- [10] B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," *Acad. Radiol.*, vol. 2, pp. 959–966, 1995.
- [11] H. D. Li, M. Kallergi, L. P. Clarke, V. K. Jain, and R. A. Clark, "Markov random field for tumor detection in digital mammography," *IEEE Trans. Med. Imag.*, vol. 14, pp. 565–576, Sept. 1995.
- [12] M. L. Giger, C. J. Vyborny, and R. A. Schmidt, "Computerized characterization of mammographic masses: Analysis of spiculation," *Cancer Lett.*, vol. 77, pp. 201–211, 1994.
- [13] T. K. Lau and W. F. Bischof, "Automated detection of breast tumors using the asymmetry approach," *Comput. Biomed. Res.*, vol. 24, no. 9, pp. 1501–1513, 1995.
- [14] M. J. Bianchi, A. Rios, and M. Kabuka, "An algorithm for detection of masses, skin contours, and enhancement of microcalcifications in mammograms," in *Proc. , Symp. Computer Assisted Radiology*, Winston-Salem, NC, June 1994, pp. 57–64.
- [15] T. Lei and W. Sewchand, "Statistical approach to x-ray CT imaging and its application in image analysis—Part II: A new stochastic model-based image segmentation technique for x-ray CT image," *IEEE Trans. Med. Imag.*, vol. 11, pp. 62–69, Feb. 1992.
- [16] Y. Wang, T. Adali, and S.-C. B. Lo, "Automatic threshold selection using histogram quantization," *SPIE J. Biomedical Optics*, vol. 2, no. 2, pp. 211–217, April 1997.
- [17] J. Zhang and J. W. Modestino, "A model-fitting approach to cluster validation with application to stochastic model-based image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 1009–1017, Oct. 1990.

- [18] H. Li, K. J. R. Liu, Y. Wang, and S. C. Lo, "Morphological filtering and stochastic modeling-based segmentation of masses on mammographic images," in *Proc. IEEE Nuclear Science Symp. Medical Imaging Conf.*, 1996, pp. 1792–1796.
- [19] J. Serra, *Image Analysis and Mathematical Morphology*. London, U. K.: Academic, 1982.
- [20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, pp. 1–38, 1977.
- [21] Y. Wang, T. Adali, C. M. Lau, and S. Y. Kung, "Quantitative analysis of MR brain image sequences by adaptive self-organizing finite mixtures," *J. VLSI Signal Processing*, vol. 18, no. 3, pp. 219–240, 1998.
- [22] Y. Wang, T. Adali, S. Y. Kung, and Z. Szabo, "Quantification and segmentation of brain tissues from MR images: A probabilistic neural network approach," *IEEE Trans. Image Processing*, vol. 7, pp. 1165–1181, Aug. 1998.
- [23] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.
- [24] J. Rissanen, "Modeling by shortest data description," *Automat.*, vol. 14, pp. 465–471, 1978.
- [25] R. A. Hummel and S. W. Zucker, "On the foundations of relaxation labeling processes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 5, pp. 267–286, Mar. 1983.
- [26] A. Hoover, G. J. Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 673–688, July 1996.
- [27] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, 1996.
- [28] A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh, "Validity-guided clustering with applications to image segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 4, pp. 112–123, May 1996.
- [29] H. Li, Y. Wang, K. J. R. Liu, S.-C. B. Lo, and M. T. Freedman, "Computerized Radiophasic Mass Detection—Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," *IEEE Trans. Med. Imag.*, vol. 20, no. 4, pp. 302–313, Apr. 2001.

Computerized Radiographic Mass Detection—Part II: Decision Support by Featured Database Visualization and Modular Neural Networks

Huai Li, Yue Wang, K. J. Ray Liu*, Shih-Chung B. Lo, and Matthew T. Freedman

Abstract—Based on the enhanced segmentation of suspicious mass areas, further development of computer-assisted mass detection may be decomposed into three distinctive machine learning tasks: 1) construction of the featured knowledge database; 2) mapping of the classified and/or unclassified data points in the database; and 3) development of an intelligent user interface. A decision support system may then be constructed as a complementary machine observer that should enhance the radiologists performance in mass detection. We adopt a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a is developed to carry out both soft and hard clustering. A visual explanation of the decision making is further invented as a decision support, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. A prototype system is developed and pilot tested to demonstrate the applicability of this framework to mammographic mass detection.

Index Terms—Feature extraction, knowledge database, mass detection, neural network, visual explanation.

I. INTRODUCTION

IN ORDER to improve mass lesion detection and classification in clinical screening and/or diagnosis of breast cancers, many sophisticated computer-assisted diagnosis (CAD) systems have been recently developed [1]–[10]. Although the clinical roles of the CAD systems may still be debatable, the fundamental role should be complementary to the radiologists’

Manuscript received February 3, 1997; revised January 9, 2001. This work was supported in part by the Department of Defense under Grants DAMD17-98-1-8045 and DAMD17-96-1-6254 through a subcontract from University of Michigan, Ann Arbor, and by the National Science Foundation (NSF) under NYI Award MIP-9457397. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. Giger. Asterisk indicates corresponding author.

H. Li is with the Electrical Engineering Department and Institute for Systems Research, University of Maryland at College Park, College Park, MD 20742 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Y. Wang is with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

*K. J. Ray Liu is with the Electrical Engineering Department and Institute for Systems Research University of Maryland at College Park, College Park, MD 20742 USA (e-mail: kjrliu@eng.umd.edu).

S.-C. B. Lo and M. T. Freedman are with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Publisher Item Identifier S 0278-0062(01)02830-0.

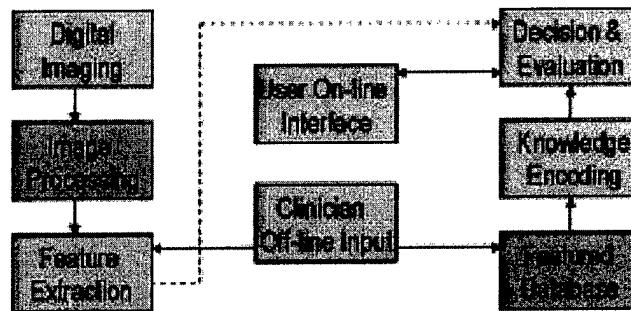


Fig. 1. Major components in CAD.

clinical duties, where the pathways of achieving ultimate performance enhancement taken by the machine observer and human observer may not necessarily be close. For example, CAD systems may attack the tasks that the radiologists cannot perform well or find difficult to perform. Because of generally larger size and complex appearance of masses, especially the existence of spicules in malignant lesions, as compared with microcalcifications, feature-based approaches are largely adopted in many CAD systems [1]–[4], [6], [7]. Kegelmeyer has first reported promising results for detecting spiculated tumors based on local edge characteristics and Laws texture features [7]. Zwiggelaar *et al.* developed a statistical model to describe and detect the abnormal pattern of linear structures of spiculated lesions [1]. Karssemeijer *et al.* [2] proposed to identify stellate distortions by using the orientation map of line-like structures. Petrick *et al.* presented to reduce the false positive detection by combining the breast tissue composition information [4]. Zhang *et al.* used the Hough spectrum to detect spiculated lesions [6].

Although many previously proposed approaches have led to impressive results [1]–[5], [7], several fundamental issues remain unresolved in the application of CAD systems. Fig. 1 shows a general block diagram of CAD systems. Previous research has demonstrated that: 1) breast cancer is missed on mammograms in part because the optical density and contrast of the cancer is not optimal for human observer; 2) computer-based detection appears to be more affected by different criteria than human perception; 3) the challenges and pathways to the human or machine observers may be quite different, and 4) decision making by the CAD systems are largely not transparent to the user. For example, the training cases contributing to the database are often selected by the human observer while the featured knowledge database is constructed through mathematical pathways of feature extraction. The mismatch

between the human supervised case selection in training and the machine dominant mass candidates selection in testing may exist. Second, the featured knowledge database is often high-dimensional with complex internal structures. Imposing a heuristically designed neural network for learning from the training data set may prevent a correct identification of the intrinsic data structure and an accurate estimation of the class boundaries. There may also exist the mismatch between the data structure and classifier architecture or between the class boundaries and decision boundaries. Furthermore, since the machine observer and human observer may not detect the same set of masses, the "black box" nature of most CAD systems to the clinical users will prevent a natural on-line integration of human intelligence and further upgrade of a CAD system. An interactive user interface should be considered to leverage the complementary roles of the CAD in the clinical practice.

As a step toward improving the performance of a CAD system, we have put considerable efforts to conduct various studies and develop reliable image enhancement and lesion selection techniques. The methods and results have been reported in [24], where the purposes of the research were to localize the potential mass sites and help accurate feature extraction. This paper addresses the further development of computer-assisted mass detection based on the 1) construction of the featured knowledge database; 2) mapping of the classified and/or unclassified data points in the database; and 3) development of an intelligent user interface (IUI). The clinical goal is to eliminate the false positive sites that correspond to normal dense tissues with *mass-like* appearances through featured discrimination. We adopt a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a probabilistic modular neural network (PMNN) is developed to carry out both soft and hard clustering. A visual explanation of the decision making is further invented as a decision support tool, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. The motivation of this work comes from the following considerations. First, though both human and machine observers use the same set of raw data in the diagnostic stage, the construction of the knowledge database for training machine classifiers and that accomplished by human brains are indeed different. Thus, the knowledge database should be established with both machine and expert organized representative cases. Second, a quantitative understanding of the knowledge database used by the machine observer should be acquired to logically compare and/or predict the performance of CAD systems with respect to the human observers without possible under- or over-estimation, and to optimize the feature extraction and design of the machine learner for best final performance. Finally, since the human and machine observers indeed take different learning and intelligence pathways, an IUI should be developed to visually (e.g., transparently) explain the entire internal decision making process of the CAD system to the human observer to enhance the clinical decision when facing either consistent or conflicting opinions.

The major differences between our work and the previous work [1]–[10] are as follows.

- 1) We construct a knowledge database by combining both expert and machine selected cases where the assignment of class memberships (e.g., mass and nonmass classes) is supervised by the radiologists or pathological report *after* all the cases are collected.
- 2) We impose a model identification procedure to determine the optimal number and kernel shape of the local clusters within each of the two classes in a high-dimensional feature space. The model is then estimated using the expectation–maximization (EM) algorithm and information theory.
- 3) We develop a PMNN, which is considered as a nonlinear classifier, to carry out the mapping function of the knowledge database. In the knowledge database, the decision likelihood boundaries and the class prior probabilities are determined in a separate fashion, and the structure of PMNN is optimized by adapting to the database structure.
- 4) We derive a probabilistic principal component projection scheme to reduce the dimensionality of the feature space for natural human perception. The scheme leads to a hierarchical visualization algorithm allowing the complete data set to be analyzed at the top level, with best separated clusters and subclusters of data points analyzed at deeper levels.

The framework of the proposed method for mass detection is illustrated in Fig. 2. A detailed description of this paper is organized as follows. In Section II, the procedure of the knowledge database construction is described. The data mapping process for decision making is presented in Section III. Section IV presents the design of the IUI for the CAD systems. Finally, major results and discussions are summarized in Section V.

II. KNOWLEDGE DATABASE CONSTRUCTION

Given the available information contained in the raw data of mass sites and in order to establish machine intelligence carried out by various machine observers, a knowledge database may be constructed in a multidimensional feature space. It should be emphasized however that the knowledge acquired by the human brain uses much more sophisticated processes than the artificial systems. Though feature extraction has been a key step in most pattern analysis tasks, the mathematical procedures are often done intuitively and heuristically. The general guidelines are:

- 1) *Discrimination*: Features of patterns in different classes should have significantly different values.
- 2) *Reliability*: Features should have similar values for the patterns of the same class.
- 3) *Independence*: Features should not be strongly correlated to each other.
- 4) *Optimality*: Some redundant features should be deleted. A small number of features is preferred for reducing the complexity of the classifier.

Many useful image features have been suggested previously by both image processing and pattern analysis communities [11]–[13]. These features can be divided into three categories, namely, intensity features, geometric features, and texture

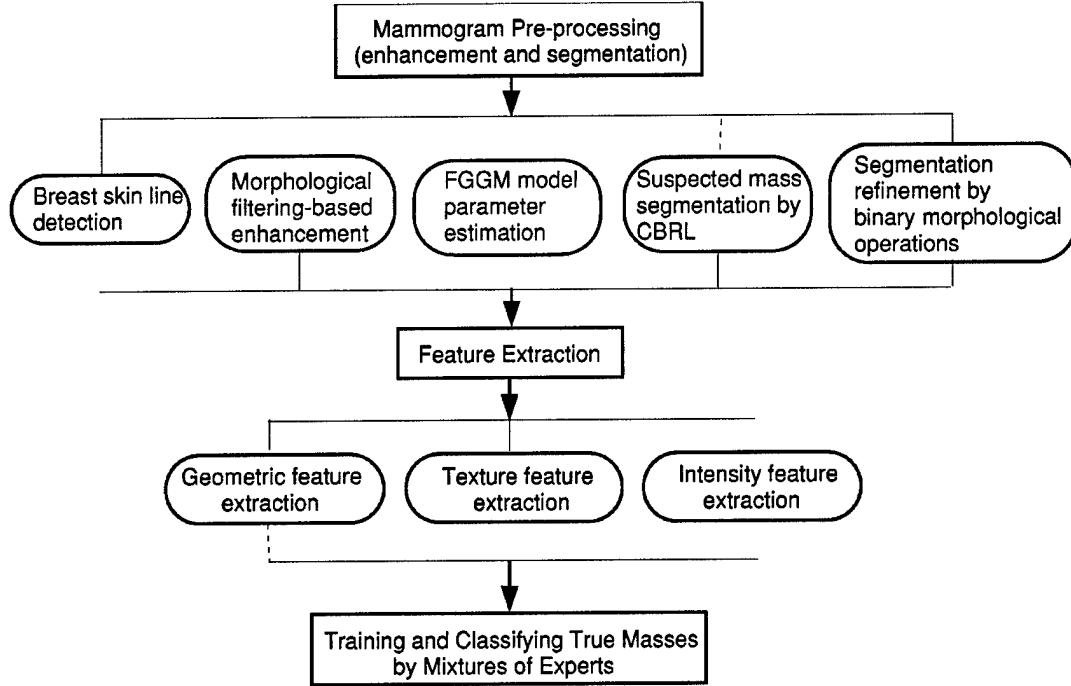


Fig. 2. The flow diagram of mass detection in digital mammograms.

features, whose values are calculated from the pixel matrices of the regions of interest (ROIs). Though these features are mathematically well defined, they may not be complete since they cannot capture all of the capable aspects of human perception nature. Thus, in this study, we have included several additional expert-suggested features to reflect the radiologists' experience. The typical features are summarized in Table I, where Fig. 3 shows the raw image of corresponding featured sites.

The joint histogram of the feature point distribution extracted from true and false mass regions are investigated, and the features that can better separate the true and false mass regions are selected for further study. Our experience has suggested that three features, i.e., the site area, two measured compactness (circularity), and difference entropy, were having better discrimination and reliability properties. Their definitions are given as follows.

1) Compactness 1

$$C_1 = \frac{A_1}{A} \quad (1)$$

where A is the area of the actual suspected region, and A_1 is the area of the overlapped region of A and the effective circle A_c , which is defined as the circle whose area is equal to A and is centered about the corresponding centroid of A .

2) Compactness 2

$$C_2 = \frac{P}{4\pi A} \quad (2)$$

where P is the boundary perimeter, and A is the area of region.

TABLE I
THE SUMMARY OF MATHEMATICAL FEATURES

Feature Sub-Space	Features
A. Intensity Features	1. contrast measure of ROIs; 2. standard derivation inside ROIs; 3. mean gradient of ROIs boundary
B. Geometric Features	1. area measure; 2. circularity measure; 3. deviation of the normalized radial length; 4. boundary roughness;
C. Texture Features	1. energy measure; 2. correlation of co-occurrence matrix; 3. inertia of co-occurrence matrix; 4. entropy of co-occurrence matrix; 5. inverse difference moment; 6. sum average; 7. sum entropy; 8. difference entropy; 9. fractal dimension of surface of ROI;

3) Difference Entropy

$$DH_{d, \theta} = - \sum_{k=0}^{L-1} p_{x-y}(k) \log p_{x-y}(k) \quad (3)$$

where

$$p_{x-y}(k) = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{d, \theta}(i, j), \quad |i - j| = k. \quad (4)$$

Several important observations are worth reiteration:

- 1) The knowledge database that will be used by the CAD system are constructed from the cases selected by both lesion localization procedure and human expert's experience. This joint set provides more complete knowledge to



Fig. 3. One example of mass segmentation and boundary extraction. (a) Mass patch; (b) segmentation; (c) boundary extraction.

the machine observer. In particular, during the interactive decision making, CAD system can still provide opinion when the cases are missed by the localization procedure but presented to the system by the radiologists.

- 2) The knowledge database is defined quantitatively in a high dimensional feature space. It provides not only the knowledge for training the machine observer, but also an objective base for evaluating the quality of feature extraction or network's learning capability, and the on-line visual explanation possibility.
- 3) The assignment of the cases' class memberships (e.g., mass and nonmass classes) is supervised by the radiologists or pathological reports. A complete knowledge database includes three subsets: raw data of mass-like sites, corresponding feature points, and class membership labels.

III. DATA MAPPING FOR DECISION MAKING

The decision making support by a CAD system addresses the problem of mapping a knowledge database, given a finite set of data examples. The mapping function can therefore be interpreted as a quantitative representation of the knowledge about the mass lesions contained in the database [14]. Instead of mapping the whole data set using a single complex network, it is more practical to design a set of simple class subnets with local mixture clusters, each one of which represents a specific region of the knowledge space. Inspired by the principle of *divide-and-conquer* in applied statistics, PMNN has become increasingly popular in machine learning research [14], [15], [19]–[22]. In this section, we present its applications to the problem of mapping from databases in mass detection, with a constructive criterion for designing the network architecture and the learning algorithm that are governed by information theory [25].

A. Statistical Modeling

The quantitative mapping of a database may be decomposed into three distinctive learning tasks: the detection of the structure of each class model with local mixture clusters; the estimation of the data distributions for each induced cluster inside each class; and the classification of the data into classes that realizes the data memberships. Recently, there has been considerable success in using finite mixture distributions data mapping [15], [17], [18], [20]. Assume that the data points \vec{x}_i in a multidimensional database come from M classes $\{\vec{\omega}_1, \dots, \vec{\omega}_r, \dots, \vec{\omega}_M\}$, and each class contains K_r clusters $\{\vec{\theta}_1, \dots, \vec{\theta}_k, \dots, \vec{\theta}_{K_r}\}$, where $\vec{\omega}_r$ is the model parameter vector of class r , and $\vec{\theta}_k$ is the kernel parameter vector of cluster k within class r . The class conditional probability measure for any data point inside the class r , i.e., the standard finite mixture distribution (SFMD), can be obtained as a sum of the following general form:

$$f(\vec{v}|\vec{\omega}_r) = \sum_{k=1}^{K_r} \pi_k g(\vec{v}|\vec{\theta}_k) \quad (5)$$

where $\pi_k = P(\vec{\theta}_k|\vec{\omega}_r)$ with a summation equal to one, and $g(\vec{v}|\vec{\theta}_k)$ is the kernel function of the local cluster distribution. For the model of global class distributions, we denote the Bayesian prior for each class by $P(\vec{\omega}_r)$. Then the sufficient statistics according to the Bayes' rule, are the posterior probability $P(\vec{\omega}_r|\vec{x}_i)$ given a particular observation \vec{x}_i

$$P(\vec{\omega}_r|\vec{x}_i) = \frac{P(\vec{\omega}_r)f(\vec{x}_i|\vec{\omega}_r)}{p(\vec{x}_i)} \quad (6)$$

where $p(\vec{x}_i) = \sum_{r=1}^M P(\vec{\omega}_r)f(\vec{x}_i|\vec{\omega}_r)$.

B. Class Distribution Learning

Class distribution learning addresses the combined estimation of regional parameters ($\pi_k, \vec{\theta}_k$) and detection of the structural parameter K_r and the kernel shape of $g(\cdot)$ in (5) based on the observations \vec{x}_r . One natural criterion used for learning the optimal parameter values is to minimize the distance between the SFMD, denoted by $f_r(\vec{u})$, and the class data histogram, denoted by $f_{x_r}(\vec{u})$ [17]. In this paper, we use relative entropy (Kullback–Leibler distance), suggested by information theory

[25], as the distance measure (for simplicity we use $f_r(\vec{u})$ to denote $f(\vec{u}|\vec{\omega}_r)$ in our formulation), given by

$$D(f_{\mathbf{x}_r} || f_r) = \sum_{\vec{u}} f_{\mathbf{x}_r}(\vec{u}) \log \frac{f_{\mathbf{x}_r}(\vec{u})}{f(\vec{u}|\vec{\omega}_r)}. \quad (7)$$

We have previously shown that when relative entropy is used as a distance measure, the distance minimization method is equivalent to the soft-split classification-based method under the criterion of maximum likelihood (ML) [23].

Another important issue concerning unsupervised distribution learning is the detection of the structural parameters of the class distribution, called model selection [15]. The objective here is to propose a systematic strategy for determining the optimal number and kernel shape of local clusters, when the prior knowledge is not available. This is indeed the case when the structure of the mass lesion patterns for a particular type of cancer may be arbitrarily complex, so correct identification of the database structure is very important. Thus, it will be desirable to have a neural network structure that is adaptive, in the sense that the number and kernel shape of local clusters are not fixed beforehand. In this paper, we applied two popular information theoretic criteria, i.e., the Akaike information criterion and minimum description length to guide the model selection procedure [24].

As the counterpart for adaptive model selection, there are many numerical techniques to perform ML estimation of cluster parameters [17]. For example, EM algorithm first calculates the posterior Bayesian probabilities of the data through the observations and the current parameter estimates (*E*-step) and then updates parameter estimates using generalized mean ergodic theorems (*M*-step). The procedure cycles back and forth between these two steps. The successive iterations increase the likelihood of the model parameters. The scheme provides winner-takes-in probability (Bayesian “soft”) splits of the data, hence allowing the data to contribute simultaneously to multiple clusters. For the sake of simplicity, we assume the kernel shape of local clusters to be a multidimensional Gaussian with mean $\vec{\mu}_{kr}$ and variance Γ_{kr} . We summarize the EM algorithm as follows.

- 1) **E-Step:** for training sample $\vec{x}^{(t)}$, $t = 1, \dots, N$, compute the probabilistic membership

$$h_{kr}^{(m)}(t) = \frac{\pi_{kr}^{(m)} p_k^{(m)}(\vec{x}^{(t)}|\vec{\omega}_r)}{\sum_{k=1}^{K_r} \pi_{kr}^{(m)} p_k^{(m)}(\vec{x}^{(t)}|\vec{\omega}_r)}. \quad (8)$$

- 2) **M-Step:** compute the updated parameter estimates

$$\pi_{kr}^{(m+1)} = \frac{1}{N} \sum_{t=1}^N h_{kr}^{(m)}(t) \quad (9)$$

$$\vec{\mu}_{kr}^{(m+1)} = \frac{1}{N\pi_{kr}^{(m+1)}} \sum_{t=1}^N h_{kr}^{(m)}(t) \vec{x}^{(t)} \quad (10)$$

$$\begin{aligned} \Gamma_{kr}^{(m+1)} &= \frac{1}{N\pi_{kr}^{(m+1)}} \sum_{t=1}^N h_{kr}^{(m)}(t) [\vec{x}^{(t)} - \vec{\mu}_{kr}^{(m+1)}] \\ &\times [\vec{x}^{(t)} - \vec{\mu}_{kr}^{(m+1)}]^T. \end{aligned} \quad (11)$$

C. Decision Boundary Learning

The objective of data classification is to realize the class membership l_{ir} for each data points based on the observation \vec{x}_i and the class statistics $\{P(\vec{\omega}_r), f(\vec{u}|\vec{\omega}_r)\}$. It is well known that the optimal data classifier is the Bayes classifier since it can achieve the minimum rate of classification error [26]. Measuring the average classification error by the mean squared error E , many previous researchers have shown that minimizing E by adjusting the parameters of class statistics is equivalent to directly approximating the posterior class probabilities when dealing with the two class problem [13], [26]. In general, for the multiple class problem the optimal Bayes classifier (minimum average error) classifies input patterns based on their posterior probabilities: input \vec{x}_i is classified to class $\vec{\omega}_r$ if

$$P(\vec{\omega}_r|\vec{x}_i) > P(\vec{\omega}_j|\vec{x}_i) \quad (12)$$

for all $j \neq r$. It should be noted that in the formulation of classifier design, the optimal criterion used for the future data classification has been intuitively and directly applied to the learning of class statistics from the training data set.

Direct learning of posterior probability is a complex task. Great effort has been made in designing the classifier as an estimator of the posterior class probability [19]. By closely investigating the global class distribution modeling, we found that the classifier design for data classification can be dramatically simplified at the learning stage. Revisit (6), since the class prior probability $P(\vec{\omega}_r)$ is a known parameter when a supervised learning is applied, the posterior class probability $P(\vec{\omega}_r|\vec{x}_i)$ can be obtained without any further effort. Thus, by conditioning $P(\vec{\omega}_r)$, the problem is formulated as a supervised classification learning of the class conditional likelihood density $f(\vec{u}|\vec{\omega}_r)$. Thus, an efficient supervised algorithm to learn the class conditional likelihood densities called the “decision-based learning” [21] is adopted in this paper. The decision-based learning algorithm uses the *misclassified* data to adjust the density functions $f(\vec{u}|\vec{\omega}_r)$, which are initially obtained using the unsupervised learning scheme described previously, so that the minimum classification error can be achieved. Define the r th class discriminant function $\phi_r(\vec{x}_i, \mathbf{w})$ to be $P(\vec{\omega}_r)f(\vec{x}_i|\vec{\omega}_r)$. Given a set of training patterns $\mathbf{X} = \{\vec{x}_i; i = 1, 2, \dots, M\}$. The set \mathbf{X} is further divided into the “positive training set” $\mathbf{X}^+ = \{\vec{x}_i; \vec{x}_i \in \vec{\omega}_r, i = 1, 2, \dots, N\}$ and the “negative training set” $\mathbf{X}^- = \{\vec{x}_i; \vec{x}_i \notin \vec{\omega}_r, i = N+1, N+2, \dots, M\}$. If the misclassified training pattern is from positive training set, reinforced learning will be applied. If the training pattern belongs to the negative training set, we anti-reinforce the learning, i.e., pull the kernels away from the problematic regions. The boundary refinement is summarized as follows:

Reinforced

$$\text{Learning: } \mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} + \eta l'(d(t)) \nabla \phi(\mathbf{x}(t), \mathbf{w})$$

Antireinforced

$$\text{Learning: } \mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} - \eta l'(d(t)) \nabla \phi(\mathbf{x}(t), \mathbf{w}) \quad (13)$$

PMNN is a probabilistic modular network designed especially for data classification where a Bayesian decomposition of the learning process provides a unique opportunity to optimize

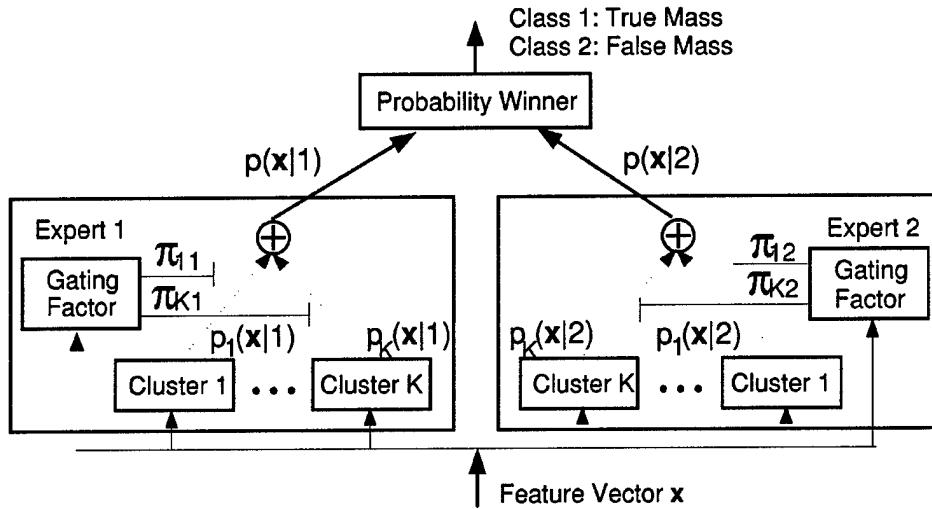


Fig. 4. The structure of the PMNN.

the structure of training scheme [14], [22]. Since the information about class population is, in general, physically uncorrelated with the conditional features about the individual class, a decoupled two-step training, in terms of both network structure and learning rule, makes much more sense than that in the conventional posterior-type neural networks, i.e., the conditional likelihood of each class and the class Bayesian prior should be adjusted separately in the classification spaces. Thus, PMNN consists of several disjoint subnets and a winner-takes-all network. The subnet outputs of the PMNN are designed to model the likelihood functions (likelihood-type network) which are first estimated from equally presented class samples, and the final decision boundaries are determined simply weighting the likelihood by the class populations. For a M -classification problem, PMNN contains M different class subnets, each of which represents one data class in the database. Within each subnet, several neurons (or clusters) are applied in order to handle problems which have complicated decision boundaries. The outputs of class subnets are fed into a winner-take-all network. The winner-take-all network categorizes the input pattern to the data class whose subnet produces the highest output value.

The structure of the PMNN used in this study is shown in Fig. 4. The PMNN consists of two subnets. Within each subnet, there are several neurons (or clusters). The outputs of class subnets are fed into a probability winner processor, which categorizes the input pattern to the data class whose subnet produces the highest probability value. The training scheme of the PMNN is based on the unsupervised learning. Each subnet is trained individually, and no mutual information across the classes may be utilized. In our study, one modular expert is trained to detect true masses, and the other is trained to detect false masses. After training, the feature vectors extracted from ROIs_{sub} are entered to this network to classify true or false masses. In both training and testing processes, we assume that the feature vectors \vec{x}_i in class r ($r = 1, \dots, M$) is a mixture of multidimensional Gaussian distributions, i.e.,

$$f(\vec{x}_i|\vec{\omega}_r) = \sum_{k=1}^{K_r} \pi_{kr} p_k(\vec{x}_i|\vec{\omega}_r) \quad (14)$$

where $\sum_{k=1}^{K_r} \pi_{kr} = 1$ and $p_k(\vec{\omega}_r) = N(\vec{\mu}_{kr}, \Gamma_{kr})$ is a multi-dimensional Gaussian distribution within cluster k of class r .

IV. INTERACTIVE VISUAL EXPLANATION

In order to improve the utility of the CAD systems in clinical practice, an IUI is highly desired. Different from many previously proposed approaches, we have organized our database from both mathematical-localized and radiologist-selected mass-like cases, and formed the featured knowledge database based on both mathematical-based and radiologist-selected image features. This off-line effort should enhance the performance of the machine observer through better quality of training set and optimal design of neural network architecture. Our experience has suggested, however, that further improvement of CAD systems requires on-line natural integration of human intelligence with the computer' output, since human perception has and can play an important role in the clinical decision making. In this research, we have pilot developed an IUI where the major functions include: 1) interactive visual explanation of the CAD decision making process; 2) on-line retrieval of the optimally displayed raw data and/or similar cases; and 3) supervised upgrade of the knowledge database by radiologist-driven input of the "unseen" and/or "typical" cases. Our preliminary studies have shown that the visual presentation of both raw data and CAD results to radiologists may provide visual cues for improved decision making.

As a step toward understanding the complex information from data and relationships, structural and discriminative knowledge reveals insight that may prove useful in data mining. Hierarchical minimax entropy modeling and probabilistic principal component projection are proposed for data explanation, which is both statistically principled and visually effective at revealing all of the interesting aspects of the data set. The methods involve multiple use of standard finite normal mixture models and probabilistic principal component projections. The strategy is that the top-level model and projection should explain the entire data set, best revealing the presence of clusters and relationships, while lower-level models and

projections should display internal structure within individual clusters, such as the presence of subclusters and attribute trends, which might not be apparent in the higher-level models and projections. With many complementary mixture models and visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet still conveys considerable structural information. In particular, a probabilistic principal component neural network is developed to generate optimal projections, leading to a hierarchical visualization algorithm. This algorithm allows the complete data set to be analyzed at the top level, with best separated subclusters of data points analyzed at deeper levels.

Research evidence suggests that for analysis of complex and high-dimensional data sets, structure decomposition and dimensionality reduction are the natural strategies in which the model-based approach and visual explanation have proven to be powerful and widely-applicable [27]. However, there is a trade-off between maximizing (structure decomposition) and minimizing (dimensionality reduction) the entropy of the system. In this research, a minimax entropy approach is adopted through the use of progressive model identification and principal component projection. The complete visual explanation hierarchy is generated by performing principal projection (dimensionality reduction) and model identification (structure decomposition) in two iterative steps using information theoretic criteria, EM algorithm, and probabilistic principal component analysis (PCA). Hierarchical probabilistic principal component visualization involves: 1) evaluation of posterior probabilities for mixture data set; 2) estimation of multiple principal component axes from probabilistic data set; and 3) generation of a complete hierarchy of visual projections.

Suppose the data space is d -dimensional with coordinates y_1, \dots, y_d and the data set consists of a set of d -dimensional vectors $\{t_i\}$ where $i = 1, \dots, N$. Now consider a three-dimensional (3-D) latent space $x = (x_1, x_2, x_3)^T$ together with a linear function which maps the latent space to the data space by $y = Wx + b$ where W is a $d \times 3$ matrix and b is a d -dimensional mean vector. If we introduce a probability distribution $p(x)$ over the latent space given by a Gaussian estimated from the latent variables $\{x_i\}$, then a similar full-dimensional Gaussian distribution in data space can be defined by convolving this distribution with a general diagonal Gaussian conditional probability distribution $p(t|x, \Lambda_d)$ in data space where Λ_d is the covariance matrix, resulting in a final form of

$$p(t) = \int p(t|x)p(x)dx \quad (15)$$

where the log likelihood function for this model is given by $L = \sum_i \log p(t_i)$. Suppose W is determined by the PCA, ML can be used to fit the model to the data and hence determine values for the parameters b and Λ_d [27]. Using a soft clustering of the data set and multiple PCAs corresponding to the clusters, a *mixture* of latent models takes the form of $p(t) = \sum_{k=1}^{K_0} \pi_k p(t|k)$ where K_0 is the number of components in the mixture, and the parameters π_k are the prior probabilities corresponding to the components $p(t|k)$. Each component is an independent latent model with PCA projection W_k and parameters b_k and Λ_{dk} . This procedure can be further extended to a hierarchical mixture model formulated by $p(t) = \sum_{k=1}^{K_0} \pi_k \sum_j \pi_{j|k} p(t|k, j)$

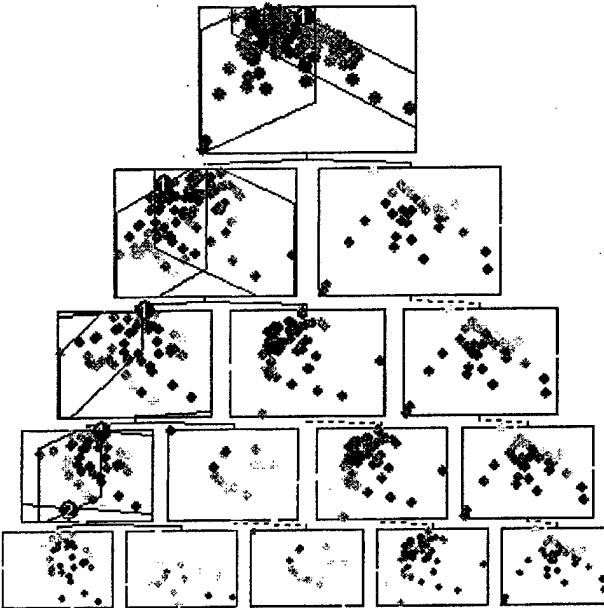


Fig. 5. The hierarchical view of computed features for mass and nonmass samples (Database A, see Table II).

where $p(t|k, j)$ again represent independent latent models [27]. With a soft partitioning of the data set via EM algorithm, data points will effectively belong to more than one cluster at any given level. This step is automatically available in our approach since the estimation of parent latent model involves the calculation of posterior probabilities denoted by π_{ik} . Thus, the effective input values are $\pi_{ik}x_i$ for an independent visualization space k , corresponding to the visualization space k in the hierarchy. It should be emphasized that *probabilistic* means both neural network based learning and posterior probability weighted inputs. Further projections can again be performed by using the effective input values $\pi_{ik}\pi_{j|k}t_i$ for the visualization subspace j . Fig. 5 shows the hierarchical view of computed features for mass and nonmass samples. In Fig. 5, a hierarchical visualization view of a high dimensional feature data set was generated using hierarchical data visualization algorithm. One hundred and 25 real cases were involved, among them 75 are mass sites, 50 are nonmass sites. Nine features were computed on 125 cases. The dimension of the resulted feature data set became 125×9 (Database A, see Table II). Hierarchical visualization tool enables the visualization of high dimensional data set through dimension reduction and data modeling so that data distribution features of the data set can be well recognized. For instance, the clusters and subclusters of mass and nonmass data points and the boundaries of the clusters can be revealed for further research purpose.

In the use of a hierarchical minimax entropy mixture model, an interactive visualization environment is required to enable a flexible computerized experiment such that a human-database interaction can be performed effectively. We have developed an interactive environment for visualizing five-dimensional (5-D) data sets, based on state-of-the-art computer graphics toolkits such as object-oriented OpenGL and OpenInventor. With a sophisticated set of various kinds of simulated lights, color

TABLE II
THE SUMMARY OF EXPERIMENTAL DATABASES

Database	Descriptions
A	Nine features extracted from 75 mass sites and 50 non-mass sites. Used for visualizing hierarchically projected high dimensional feature space. Result is presented in Figure 5.
B	A simulated two-dimensional feature space. Used to show the effect of model selection on decision boundary estimation. Result is shown in Figure 6.
C	ORL standard database. Used to show the improvement of PMNN with decision-based learning. Result is discussed in the text.
D	The training data set consisting of 50 mammograms, with 50 true mass sites and 50 false mass sites. Three most discriminatory features are extracted. Used for both PMNN training and visualization. Result is given in Figure 7.
E	The testing data set consisting of 46 mammograms, with 23 normal cases and 23 biopsy proven mass cases with each of them having at least one true mass site. Three most discriminatory features, the same as database D, are extracted. Used to test the overall performance of our CAD system prototype where the mass candidates were selected using the method reported in Part I, automatically. Result is shown in Figure 8 and also discussed in the text.

texturing editors, and 3-D manipulator and viewers (we have integrated 3-D mouse and stereo glass units into our existing system), our system allows one to examine the volumetric data sets with any viewpoint and dynamically walk through its internal structures to better understand the spatial relationships among clusters and decision surfaces present. One of the most important features in our approach is to attach the decision surface to the 3-D probability cloud in support of decision making, and to link each data point in the visualization space to its raw data so that the user can on-line retrieve the corresponding raw data such as an original image for interim decision making.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we present the experimental results using the information theoretic criteria and PMNNs to generate the mapping function of the featured database, and the preliminary results using the hierarchical minimax entropy projections to conduct visual explanation of the decision making. For the validation of the database mapping using the proposed algorithms, global relative entropy (GRE) value between the (SFMD) and the joint histogram is used as an objective measure to evaluate the fitness of the mapping function. A summary of the databases we used in our study is presented in Table II.

As we have discussed in Sections III and IV, model selection is the first and a very important learning task in mapping a database and the objective of the procedure is to determine both the number and the kernel shape of local clusters in each class. This procedure is used not only in the data mapping for decision making but also in the structure decomposition for hierarchical visual explanation. Our experience has suggested that an incorrect model selection will affect the performance of data-classification based decision making. For the sake of simplicity, we discuss this conclusion in the following 2-D example. Let us form a simulated featured database with two major features that well characterize the two targeted classes, as it shown in Fig. 6 (Database B, see Table II). The ground truth is that class 1 contains only one local cluster while class 2 contains two local clusters. With a model selection procedure

using the proposed criteria, the intrinsic data structure was correctly identified. According to the principle of designing the optimal structure of PMNN and visual explanation hierarchy, the result of these criteria also determines the most appropriate number of mixture components in the corresponding PMNN and projected cluster decomposition. Two PMNN with different architecture orders were designed and trained to determine the classification boundaries between the two classes. The classification results are shown in Fig. 6(a) and (b). The result in Fig. 6(a) is with the right cluster number in Class 2, while the result in Fig. 6(b) is with the wrong cluster number in Class 2. From this simple experiment, we have shown that the decision boundary with the right cluster number may be much more accurate than that with heuristically determined cluster number, since the decision boundary between class 1 and class 2 will be determined by four cross points in the first case while in the second case the decision boundary will be determined by only two cross points. It should be emphasized that the error of data classification is theoretically controlled by the accuracy in estimating the decision boundaries between classes, and the quality of the boundary estimates is indeed dependent upon the correct structure of the class likelihood function.

As we have discussed before, although the knowledge database contains both machine-localized and human-selected cases, in clinical settings "unseen" and/or subtle cases contribute the major false positives. We have also pilot tested the PMNN method to the so-called " $M + 1$ classes" problem, in which the disease pattern under testing could be either from one of the M classes, or from some other unknown classes (the "unknown" class or the "intruder" class). Note that the unknown class probability is often very hard to estimate because of the lack of sufficient training samples (for example, in the mass detection problem, the unknown classes include the ROIsub over the normal tissues). In our experiment, PMNN uses different decision rule from that of the " M classes" problem: pattern \vec{x}_i belongs to class r if both of the following conditions are true: a) $\phi(\vec{w}_r, \vec{x}_i) > \phi(\vec{w}_j, \vec{x}_i)$, $\forall j \neq r$, and b) $\phi(\vec{w}_r, \vec{x}_i) > T$. T is a threshold obtained by decision-based

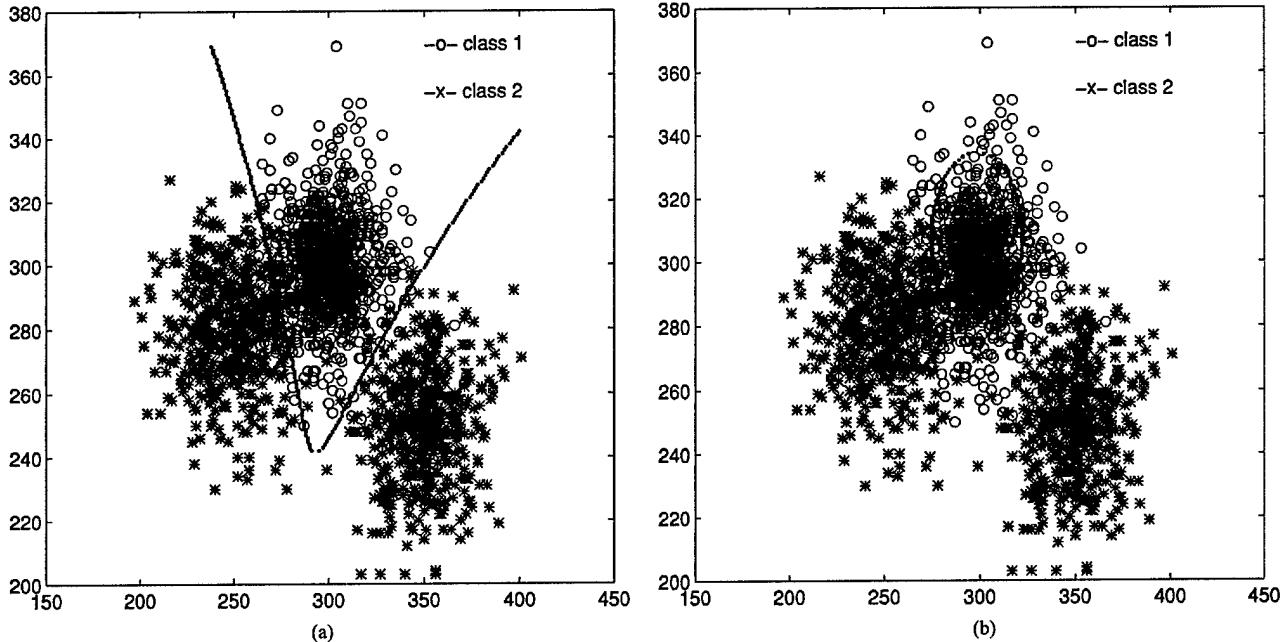


Fig. 6. The classification examples with a two-dimensional (2-D) simulated database (Database B, see Table II). (a) Class 2 contains two local clusters. (b) Class 2 contains one local cluster.

learning. Otherwise pattern \vec{x}_i belongs to the unknown class. We observed consistent and significant improvement in classification results compared with the pure Bayesian decision. Using the ORL (Olivetti Research Laboratory, Cambridge, U.K.) standard database (Database C, see Table II), our experience has shown an increase of correct detection rate from 70% to 90% [14].

In the third experiment, we use the proposed classifier to distinguish true masses from false masses based on the features extracted from the suspected regions. The objective is to reduce the number of suspicious regions and identify the true masses. 150 mammograms, each of them contains at least one mass case of varying size and location, were selected in our study. The areas of suspicious masses were identified following the proposed procedure with biopsy proven results. Fifty mammograms with biopsy proven masses were selected from the 150 mammograms for training (Database D, see Table II). The mammogram set used for testing contained 46 single-view mammograms: 23 normal cases and 23 with biopsy proven masses (Database E, see Table II) which were also selected from the 150 mammograms. All mammograms were digitized with an image resolution of $100 \mu\text{m} \times 100 \mu\text{m}/\text{pixel}$ by the laser film digitizer (Model: Lumiscan 150). The image sizes are $1792 \times 2560 \times 12 \text{ bpp}$. For this study, we shrunk the digital mammograms with the resolution of $400 \mu\text{m}$ by averaging 4×4 pixels into one pixel. According to radiologists, the size of the small masses is 3–15 mm. The middle size of masses is 15–30 mm. The large size of masses is 30–50 mm, which are rare in mammograms. A 3-mm object in an original mammogram occupies 30 pixels in a digitized image with a $100-\mu\text{m}$ resolution. After reducing the image size by four times, the object will occupy the range of about seven to eight pixels. The object with the size of seven pixels is expected to be detectable by any computer algorithm.

Therefore, the shrinking step is applicable for mass cases and can save computation time.

After the segmentation, the area index feature was first used to eliminate the nonmass regions. In our study, we set $A_1 = 7 \times 7$ pixels and $A_2 = 75 \times 75$ pixels as the thresholds. A_1 corresponds to the smallest size of masses (3 mm), and an object with a area of 75×75 pixels corresponds to 30 mm in the original mammogram. This indicates that the scheme can detect all masses with sizes up to 30 mm. Masses larger than 30 mm are rare cases in the clinical setting. When the segmented region satisfied the condition $A_1 \leq A \leq A_2$, the region was considered to be suspicious for mass. For the purpose of representative demonstration, we have selected a 3-D feature space consisting of compactness I, compactness II, and difference entropy. According to our investigation, these three features have the better separation (discrimination) between the true and false mass classes. It should be noticed that the feature vector can easily extend to higher dimensionality. A training feature vector set was constructed from 50 true mass ROIsub and 50 false mass ROIsub (Database D, see Table II). The training set was used to train two modular probabilistic decision-based neural networks separately. In addition to the decision boundaries recommended by the computer algorithms, a visual explanation interface has also been integrated with 3-D to 2-D hierarchical projections. Fig. 7(a) shows the database map projection with compactness definition I and difference entropy. Fig. 7(b) shows the database map projection with compactness definition II and difference entropy. Our experience has suggested that the recognition rate with compactness I are more reliable than that with compactness II. In order to have more accurate texture information, the computation of the second-order joint probability matrix $p_{d,\theta}(i,j)$ is only based on the segmented region of the original mammogram. For the shrunk mammograms, we found that

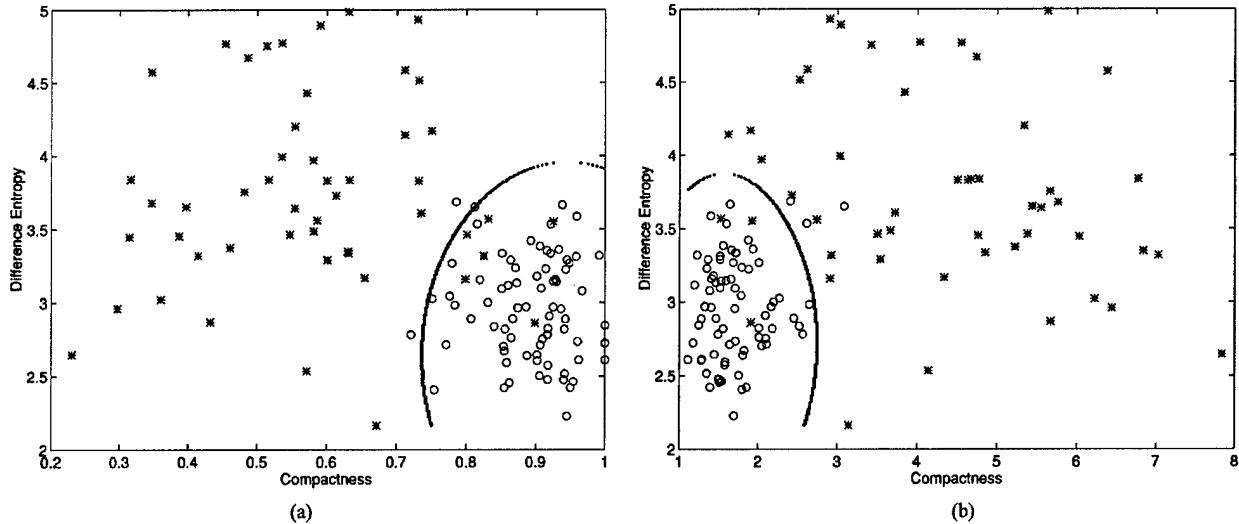


Fig. 7. The data mapping results (Database D, see Table II). -o- denotes true mass cases; -*- denotes false mass cases. (a) The mapping using compactness I. (b) The mapping using compactness II.

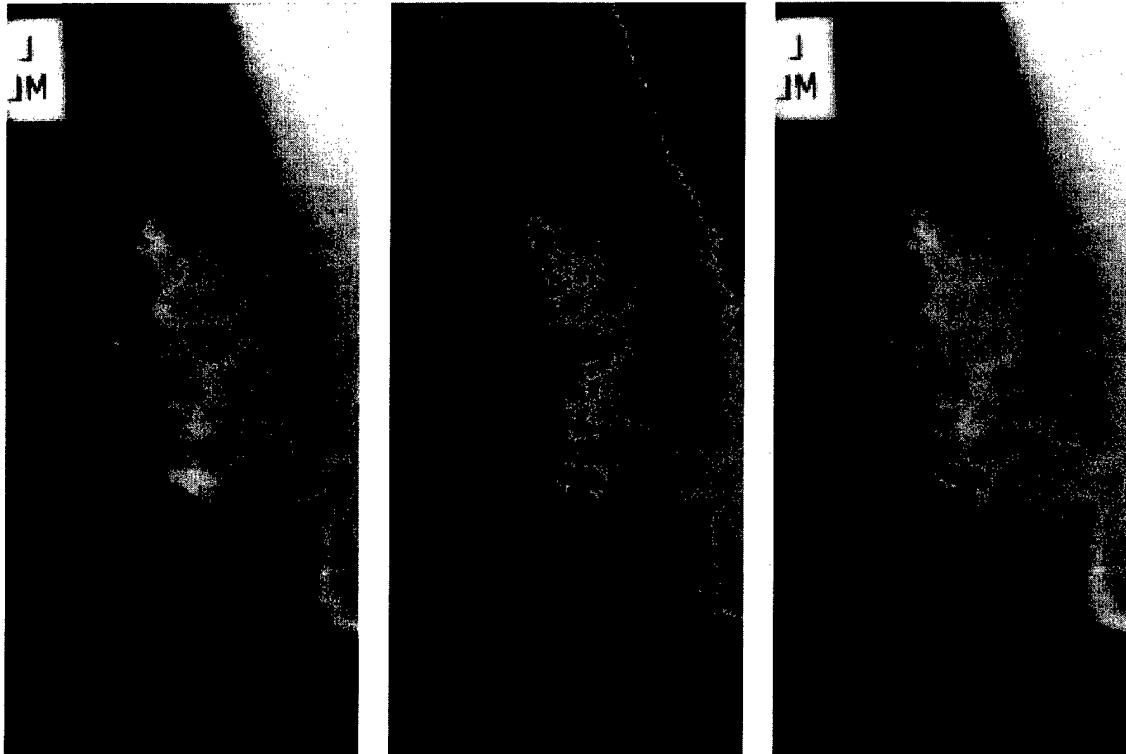


Fig. 8. One example of the mass detection using the proposed approach (Database E, see Table II).

the difference entropy had better discrimination with $d = 1$. The difference entropy used in this study was the average of values at $\theta = 0^\circ, 45^\circ, 90^\circ$, and 135° .

We have conducted a preliminary study to evaluate the performance of the algorithms in real case detection, in which 6–15 suspected masses/mammogram were detected and required further clinical decision making. We found that the proposed classifier can reduce the number of suspicious masses with a sensitivity of 84% at 1.6 false positive findings/mammogram based on the testing data set containing 46 mammograms (23 of them

have biopsy proven masses) (Database E, see Table II). Fig. 8 shows a representative mass detection result on one mammogram with a stellate mass. After the enhancement, ten regions with brightest intensity were segmented. Using the area criterion, too large and too small regions were eliminated first and the rest regions were submitted to the PMNN for further evaluation. The results indicated that the stellate mass lesion was correctly detected.

For further evaluation, receiver operating characteristic (ROC) method may be employed. However, we do not feel

ROC analysis will provide really a better evaluation but an alternative method to this case. First, most ROC analysis reported by others were based on different database thus are not comparable since ROC results are highly data-dependent. Second, ROC analysis only indicate an "overall" performance with limitations at least in twofold: it is for multithreshold thus the corresponding system may not be optimal to a particular application where only one threshold is needed; and it cannot provide a mathematically traceable feedback to improve the performance of the system or the one component in the system. Third, currently used FROC analysis package imposes several assumptions on the distributions of the cases which are invalid in most applications and particularly untrue in our situation. For example, our assumptions about the data distributions is SFNM that is clearly different from the restricted conditions imposed by the application of existing FROC analysis algorithm. In our approach, a quantitative mapping of the knowledge database is performed with hierarchical SFMD modeling and should be perfectly (at least in the theoretical sense) carried out by the corresponding PMNN classifier. In other words, optimal decision making should have already been achieved according to the Bayesian rule. It is reasonable to acknowledge that in order to compare the overall performance with the other systems, an ROC study may be further conducted. We are currently working on developing a new generation of FROC analysis package with a caution to remove the forementioned problems.

Another important consideration with the present approach is the measure of quality in visual explanation [29]. This is not a glamorous area, but progress in this area is eminently critical to the future success of visual exploration [28]. What is the correct matrix for a direct projection of a particular multimodal data set? How effective was a particular visualization tool? Did the user come to the correct conclusion? It may be agreeable that the benchmark criteria in visual exploration are very different and difficult [28]. As shared by Bishop and Tipping [27], we believe that in data visualization there is no objective measure of quality, and so it is difficult to quantify the merit of a particular data visualization technique, and the effectiveness of such a techniques is often highly data-dependent. The possible alternative is to perform a rigorous psychological evaluation using simple and controlled environment, or to invite domain experts to direct evaluate the efficacy of the algorithm for a specified task. For example, we can compare the domain expert's performances with and without the system aid. In that case, the ROC method may be used to evaluate the performance of our algorithm when used by the radiologists. While the optimality of these new techniques is often highly data-dependent, we would expect the hierarchical visualization model to be a very effective tool for the data visualization and exploration in many applications.

In summary, we employed a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points was then obtained by learning the generalized normal mixtures and decision boundaries. A visual explanation of the decision making was further invented as a decision support, based on an interactive

visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. A prototype system was developed and pilot tested to demonstrate the applicability of this framework to mammographic mass detection.

ACKNOWLEDGMENT

The authors would like to thank R. F. Wagner of the Food and Drug Administration and S.-Y. Kung of the Princeton University for their valuable scientific input.

REFERENCES

- [1] R. Zwigelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. M. Boggis, "Model-based detection of spiculated lesions in mammograms," *Med. Image Anal.*, vol. 3, no. 1, pp. 39-62, 1999.
- [2] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammogram," *IEEE Trans. Med. Imag.*, vol. 15, pp. 611-619, Oct. 1996.
- [3] L. Miller and N. Ramsey, "The detection of malignant masses by nonlinear multiscale analysis," *Excerpta Medica*, vol. 1119, pp. 335-340, 1996.
- [4] N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, M. M. Goodlett, and D. D. Adler, "Computer-aided breast mass detection: False positive reduction using breast tissue composition," *Excerpta Medica*, vol. 1119, pp. 373-378, 1996.
- [5] W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms," *Excerpta Medica*, vol. 1119, pp. 411-415, 1996.
- [6] M. Zhang, M. L. Giger, C. J. Vyborny, and K. Doi, "Mammographic texture analysis for the detection of spiculated lesions," *Excerpta Medica*, vol. 1119, pp. 347-351, 1996.
- [7] W. P. Kegelmeyer Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," *Radiology*, vol. 191, pp. 331-337, 1994.
- [8] R. N. Strickland, "Tumor detection in nonstationary backgrounds," *IEEE Trans. Med. Imag.*, vol. 13, pp. 491-499, June 1994.
- [9] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Alder, M. M. Goodlett, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857-876, 1995.
- [10] M. L. Giger, C. J. Vyborny, and R. A. Schmidt, "Computerized characterization of mammographic masses: Analysis of spiculation," *Cancer Lett.*, vol. 77, pp. 201-211, 1994.
- [11] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [12] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610-621, Nov. 1973.
- [13] R. Schalkoff, *Pattern Recognition: Statistical, Structural, and Neural Approaches*. New York: Wiley, 1992.
- [14] Y. Wang, S. H. Lin, H. Li, and S. Y. Kung, "Data mapping by probabilistic modular networks and information theoretic criteria," *IEEE Trans. Signal Processing*, vol. 46, pp. 3378-3397, Dec. 1998.
- [15] L. Perlovsky and M. McManus, "Maximum likelihood neural networks for sensor fusion and adaptive classification," *Neural Networks*, vol. 4, pp. 89-102, 1991.
- [16] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1990, pp. 1361-1364.
- [17] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.
- [18] C. E. Priebe, "Adaptive mixtures," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 910-912, 1994.
- [19] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: MacMillan College, 1994.
- [20] M. I. Jordan and R. A. Jacobs, "Hierarchical mixture of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181-214, 1994.
- [21] S. Y. Kung and J. S. Taur, "Decision-based neural networks with signal/image classification applications," *IEEE Trans. Neural Networks*, vol. 1, pp. 170-181, Jan. 1995.

- [22] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE Trans. Neural Networks (Special issue on Artificial Neural Networks and Pattern Recognition)*, vol. 8, Jan. 1997.
- [23] Y. Wang, L. Luo, H. Li, and M. T. Freedman, "Hierarchical minimax entropy modeling and probabilistic principal component visualization for data explanation and exploration," presented at the SPIE Medical Imaging Conf., San Diego, CA, Feb. 20–26, 1999.
- [24] H. Li, Y. Wang, K. J. R. Liu, S.-C. B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection—Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," *IEEE Trans. Med. Imag.*, vol. 20, no. 4, pp. 289–301, Apr. 2001.
- [25] T. W. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [26] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Berlin, Germany: Springer-Verlag, 1988.
- [27] C. M. Bishop and M. E. Tipping, "A hierarchical latent variable model for data visualization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 281–293, Mar. 1998.
- [28] G. M. Nielson, "Challenges in visualization research," *IEEE Trans. Visual. Comput. Graphics*, vol. 2, pp. 97–99, 1996.
- [29] E. R. Tufte, *Visual Explanation: Images and Quantities, Evidence and Narrative*. Cheshire, U.K.: Graphics, 1996.

A Multiple Circular Path Convolution Neural Network System for Detection of Mammographic Masses

Shih-Chung B. Lo*, Member, IEEE, Huai Li, Member, IEEE, Yue Wang, Member, IEEE, Lisa Kinnard, and Matthew T. Freedman

Abstract—A multiple circular path convolution neural network (MCPCNN) architecture specifically designed for the analysis of tumor and tumor-like structures has been constructed. We first divided each suspected tumor area into sectors and computed the defined mass features for each sector independently. These sector features were used on the input layer and were coordinated by convolution kernels of different sizes that propagated signals to the second layer in the neural network system. The convolution kernels were trained, as required, by presenting the training cases to the neural network.

In this study, randomly selected mammograms were processed by a dual morphological enhancement technique. Radiodense areas were isolated and were delineated using a region growing algorithm. The boundary of each region of interest was then divided into 36 sectors using 36 equi-angular dividers radiated from the center of the region. A total of 144 Breast Imaging—Reporting and Data System-based features (i.e., four features per sector for 36 sectors) were computed as input values for the evaluation of this newly invented neural network system. The overall performance was 0.78–0.80 for the areas (A_x) under the receiver operating characteristic curves using the conventional feed-forward neural network in the detection of mammographic masses. The performance was markedly improved with A_x values ranging from 0.84 to 0.89 using the MCPCNN. This paper does not intend to claim the best mass detection system. Instead it reports a potentially better neural network structure for analyzing a set of the mass features defined by an investigator.

Manuscript received February 22, 2000; revised January 11, 2002. This work was supported by the US Army under Grant DAMD17-96-1-6254 through a subcontract from University of Michigan, Ann Arbor, and under Grant DAMD17-01-1-0267 through a subcontract from Howard University. The work of Y. Wang was supported by the US Army under Grant DAMD17-98-1-8045. The work of L. Kinnard was supported by the US Army under Grant DAMD17-00-1-0291. The content of this paper does not necessarily reflect the position or policy of the government. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. Asterisk indicates corresponding author.

*S.-C. B. Lo is with the Center for Imaging Science and Information System, Radiology Department, Georgetown University Medical Center, 2115 Wisconsin Avenue, Suite 603, N.W., Washington, DC 20007 USA (e-mail: lo@isis.georgetown.edu).

H. Li was with the ISIS Center, Radiology Department, Georgetown University Medical Center, Washington, DC 20007 USA. He is now with the Center for Information Technology, Division of Computational Bioscience, National Institutes of Health, Bethesda, MD 20892 USA.

Y. Wang is with the Department of Electrical Engineering and Computer Sciences, The Catholic University of America, Washington, DC 20064 USA.

L. Kinnard is with the Center for Imaging Science and Information System, Radiology Department, Georgetown University Medical Center, Washington, DC 20007 USA, and also with the Department of Electrical Engineering, Howard University, Washington, DC 20059 USA.

M. T. Freedman is with the Center for Imaging Science and Information System, Radiology Department, Georgetown University Medical Center, Washington, DC 20007 USA.

Publisher Item Identifier S 0278-0062(02)02935-X.

Index Terms—BI-RAD, computer-aided diagnosis, convolution neural network, mammography masses, neural network, sector features.

I. INTRODUCTION

IT IS KNOWN that effective treatment of breast cancer calls for early detection of cancerous lesions (e.g., clustered microcalcifications and masses associated with malignant cellular processes) [1]–[3]. Breast masses appear as areas of increased density on mammograms. It is particularly difficult for radiologists to detect and analyze a suspected area where a mass is overlapped with dense breast tissue. These masses are more readily seen as time progresses, but the further the tumor has progressed, the lower the possibility of a successful treatment. Therefore, increasing the chances of early breast cancer detection in improving today's clinical system is of vital importance in breast cancer diagnosis.

Several research groups have developed computer algorithms for automated detection of mammographic masses [4]–[8]. Some of these methods involved in classification of masses and normal dense breast tissues [7], [8]. Investigators also attempted to classify the malignant or benign nature of the detected tumors [9]–[11]. It is conceivable that correct segmentation of the masses [12] plays an important processing step prior to further mass analysis. In short, the results of these detection programs indicate that a high true-positive (TP) rate can be obtained at the expense of two or three false-positive (FP) detections per mammogram. Mammographically, a multiplicity (more than two) of similar benign-appearing breast lesions argues strongly for benignity [13]–[16] and, indeed, the more masses that are identified, the less chance that they represent cancer [17]. If the computer indicates multiple suspicious locations on a mammogram, the radiologist has to seek out one mass that possesses mammographic features, which are different from the others. The significant lesion may be missed due to the multiplicity of possible lesions. We, therefore, believe that a more useful and fundamental approach to computer-aided diagnosis (CAD) of masses is to devise computer programs to analyze features of a suspected area [18], [19] and to provide feature measures and estimates of the likelihood of malignancy by making comparisons within a digital mammographic database. The computer, therefore, serves as a second opinion and also provides a reproducible and an objective evaluation of the mass. With this aid, the radiologist may also increase his/her sensitivity by lowering the threshold of suspicion, while maintaining the overall specificity and reading efficiency.

II. CLINICAL BACKGROUND OF BREAST LESIONS AND TECHNICAL APPROACH IN MASS DETECTION

A. Description of Clinical Background

Most commonly, breast cancer presents itself as a mass. The same lesion shows a somewhat different picture from one projection to the other. Difficulties in masses also vary with the underlying breast parenchyma. In the fatty breast, masses are generally easy to detect. In the dense breast, mass detection is more difficult and auxiliary signs aid this detection. When the breast contains one mass, the decision process is based on its size, shape, and margins. When there are several masses, one looks at each, trying to determine whether any has features to suggest cancer. Furthermore, one looks to see if any mass is different in appearance from the others. Multiple small, well-defined, similar masses that present themselves bilaterally are all likely to be benign. Large, poorly defined, spiculated and unusually radiodense masses are extremely likely to be malignant. In this study, we used several computational features (see Section III-B) highly associated with four major features of breast masses routinely used in clinical reading:

Density: Malignant lesions tend to have greater radiographic density due to high attenuation and less compressibility of cancer than normal tissue. Radiolucent lesions are typically benign and the diagnosis can be made from the mammogram.

Size: If the lesion has morphological features suggesting malignancy, it should be considered suspicious regardless of the size. Isolated masses with noncystic densities greater than 8 mm in diameter can be malignant. In general, the larger a lesion, the more suspicious it is.

Shape: The more irregular the shape of a lesion, the more likely the possibility of malignancy. Lesions tend to be round, ovoid and/or lobulated. Small and frequent lobulations are suspicious. Lesions in the lateral aspect of the breast near the edge of the parenchyma with a reniform shape and a hilar indentation or notch usually represent a benign intramammary lymph node. Breast carcinoma hidden in the dense tissues can cause parenchymal retraction, which possess different shapes.

Margins: The margins of the lesion should be carefully evaluated for areas of spiculation, stellate patterns or ill-defined regions. Most breast cancers have ill-defined margins secondary to tumor infiltration and associated fibrosis. The appearance of spiculations and a more diffuse stellate pattern are almost pathognomonic for cancer. Lesions with sharply defined margins have a high likelihood of being benign; however, up to 7% of malignant lesions can be well circumscribed.

These are known clinical features and have been adapted in "Breast Imaging—Reporting and Data System" (BI-RAD) [20] of the American College of Radiology. Fig. 1(a) and (b) shows two breast images containing masses. In Fig. 1(a), a malignant mass is superimposed on the dense glandular tissue.

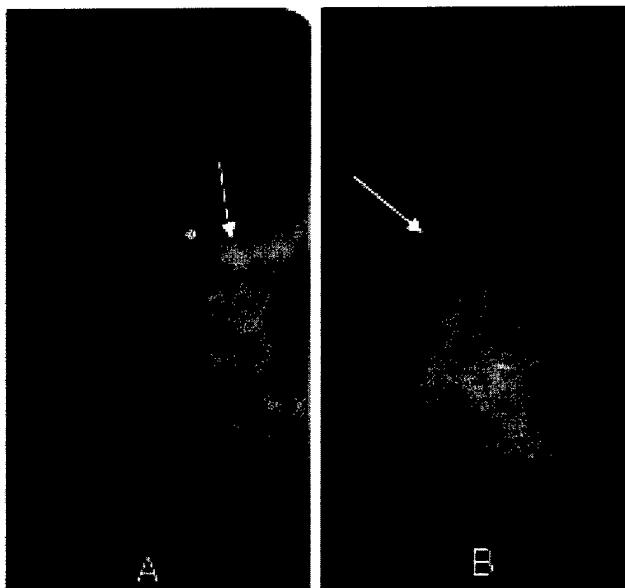


Fig. 1. (a) Dense breast containing a malignant mass. (b) Fatty and glandular breast containing a malignant mass.

However, its spiculated nature makes it easily identifiable. In Fig. 1(b), another malignant mass is located on the fatty background but is associated with a large body of glandular tissue. This mass is not easily detectable by the computer because its density is lower than the neighboring glandular tissue. Furthermore, one end of the mass is fully connected with this tissue.

B. Technical Approach for Detection of Mammographic Masses

In this study, our goal was to detect clinically suspicious lesions. The differentiation of benign and malignant status of the mammographic masses can be extended from this study model and will be reported in our future work. The study was conducted with the following steps: 1) use background correction method and morphological operations to extract radio-opaque areas; 2) delineate the boundary of the areas; 3) compute the features and texture of the masses with emphasis on the boundary; and 4) design training strategy using neural networks as classifiers for the recognition of mass features. The overall detection scheme of the study framework is shown in Fig. 2.

III. DEVELOPMENT OF TECHNICAL METHODS

A. Preprocessing and Extraction of Suspicious Masses

In automatic mass detection, accurate selection of suspected masses is considered a critical first step due to the variability of normal breast tissue and the lower contrast and ill-defined margins of masses. In our previous study [18], we aimed to improve the task of lesion site selection using model-based image processing techniques for unsupervised lesion site selection. We focused on two essential issues in the stochastic model-based image segmentation: enhancement and model selection. Based on the differential geometric characteristics of masses against

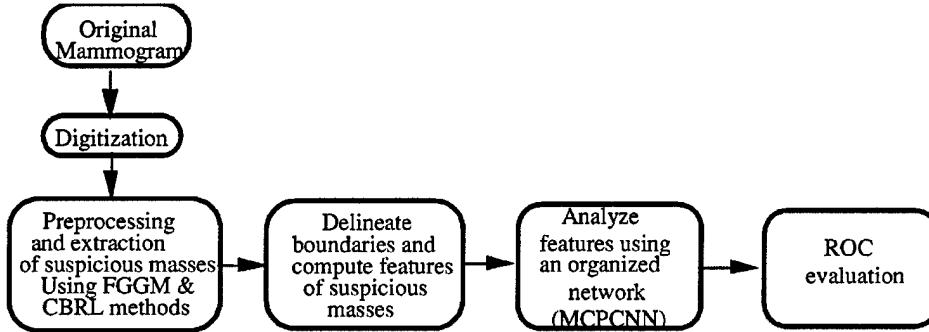


Fig. 2. A system flow chart for the detection of masses in this study.

the background tissues, we proposed one type of morphological operation to enhance the mass patterns on mammograms by removing high intensity background caused by breast tissues while maintaining mass-signals [18]. Then we employed a finite generalized Gaussian mixture (FGGM) distribution to model the histogram of the mammograms where the statistical properties of the pixel images are largely unknown and are to be incorporated. We incorporate the expectation-maximization algorithm with two information theoretic criteria to determine the optimal number of image regions and the kernel shape in the FGGM model. Finally, we applied a contextual Bayesian relaxation labeling (CBRL) technique to perform the selection of suspected masses.

We consistently processed the mammograms using this prescreening segmentation method. In the previous study [18], the FGGM method isolated 1142 potential masses including 114 of the 186 true masses in 200 mammograms. The mammograms were collected from the Mammographic Image Analysis Society (MIAS) database [21] and Brook Army Medical Center (BAMC) database. After morphological enhancement, 3143 potential masses were extracted using the FGGM technique. Of them, 181 were masses; however, five masses were not extracted. The results demonstrated that more true masses were picked up after enhancement although more false cases were also included. The undetected areas mainly occurred at the lower intensity side of the shaded objects or more obscured by fibroglandular tissues that, however, were extracted on morphological enhanced mammograms. Additionally, when the margins of masses are ill defined, only parts of suspicious masses were extracted from the original mammograms. We, therefore, decided to use the proposed morphological operation as a preprocessing step for the image enhancement prior to a segmentation method for the extraction of potential masses on the mammograms.

Based on the CBRL segmented region of interest (ROI), we employed a region growing method using a four-neighbors connection method assisted with a template masking operation to fill unconnected holes in the ROI

$$\text{IF } f(x-a, y-b) > V \text{ and } f(x, y) \in S, \\ \text{then } f(x-a, y-b) \in S \quad (1)$$

$$\text{IF } f(x-d, y-d) \in S, \text{ then } f(x-t, y-s) \in S \\ \text{for } t \leq d \text{ and } s \leq d \quad (2)$$

where V denotes the threshold value of the originally CBRL segmented ROI, S represents the set of growing region, and $[a, b]$ is a set of four conditions (i.e., $[1, 0]$, $[-1, 0]$, $[0, 1]$, and $[0, -1]$) for the four neighboring pixels. In (2), d is the size of template. In practice, we found that d should be set at five pixels to fill the holes without disrupting the boundary.

B. Feature Extraction of the Masses

Feature extraction methods play an essential role in many pattern recognition tasks. Once the features associated with an image pattern are extracted accurately, they can be used to distinguish one class of patterns from the others. Recently, many investigators have found that the multilayer perceptron (MLP) neural network using the error backpropagation training technique is a very powerful tool to serve as a classifier [22], [23]. In fact, the use of MLP neural network system for classification of disease patterns has been widely applied in the field of CAD [24]–[28].

The success of using a classifier for a pattern recognition task would rely on two factors: 1) selected features that could describe a discrepancy between image patterns and 2) accuracy of the feature computation. Should either one fail, no analyzer or classifier would be able to achieve an expected performance. By analyzing many clinical samples of various sizes of masses, we found that the peripheral portion of the mass plays an important role for mammographers to make a diagnosis. The mammographer usually evaluates the surrounding background of a radio-dense area when a region is suspected.

We used the CBRL segmented ROI to compute the center. Since the segmented ROIs were somewhat smaller than the mammographer's delineation and on the denser region of the suspected patch, the computed centers were quite close to the visual center. We then divided the boundary of the ROI into 36 sectors (i.e., 10° per sector) using 36 equi-angular dividers radiated from the center of the ROI. The following features were computed within each 10° sector of the region.

- a) “ l ”—the length from the center of the ROI to the boundary segment of the sector.
- b) “ a ”—the $\cos(\theta)$ (where θ is the normal angle of the boundary).
- c) “ g ”—the average gradient of gray value on the segment along the radial direction (i.e., $g = \sum_{i=1}^N \{g_i/N\}$) where N is the number of pixels of i along the radial direction from $l/3$ inside the boundary to the boundary (see the left

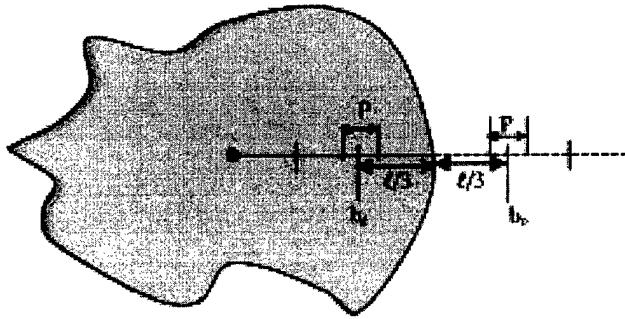


Fig. 3. A suspicious mass is delineated and shown as the shaded region. Contrast is computed by subtracting the average background pixel value (i.e., b_o , $o = 1, 2, \dots, P$) from the average foreground value (i.e., h_i , $i = 1, 2, \dots, P$).

$l/3$ line segment, Fig. 3). Technically speaking, this set of gradient values may also serve as a fuzzy system on the input layer in the neural network (to be described in Section III-C).

- d) “ c ”—the gray value difference (i.e., contrast) along the radial direction. Specifically, $c = \sum_{i=1}^P \{h_i/P\} - \sum_{o=1}^P \{b_o/P\}$ where h_i (or b_o) represents a pixel value along the radial direction. The position $l/3$ inside the boundary is the center of pixels h_i ($i = 1, 2, 3, \dots, P$) and position $l/3$ outside the boundary is the center of pixels b_o ($o = 1, 2, 3, \dots, P$), and P is the number of pixels equivalent to a segment of $l/6$ and was used for averaging (see Fig. 3).

Hence, a total of 144 computed features (four features/sector for 36 sectors) were used as input values for the classification of the ROI. The relationship between the computed features and BI—RADS descriptors are discussed below.

- i) ROI Size—The size of ROI is provided by the 36 “ l ” values.
- ii) ROI Shape (round, oval, lobulated, or irregular)—The 36 “ l ” and 36 “ a ” values can describe the shape of the ROI.
- iii) ROI Margin (circumscribed, microlobulated, obscured, ill-defined, or spiculate)—The 36 “ g ” and 36 “ l ” values can describe the ROI margin.
- iv) ROI Density (fat-containing, low density, isodense, or highly dense)—The 36 “ c ” and 36 “ g ” values can be used to describe the density distribution of the ROI.

In short, the selected features are greatly associated with the main mass descriptors indicated in the BI—RADS. The reason for using 36 values for each nominated feature is four-fold: 1) mass boundary varies, it is difficult to describe an image pattern using a single value; 2) due to the general shape of the masses, the features of masses can be easily analyzed by the polar coordinate system; 3) in case some features are inaccurately computed in several directions due to the structure noises, such as the breast slender lines, there may still exist a sufficient number of correct features; and 4) generally more accurate results can be produced by using subdivided parameters rather than using global parameters in a pattern recognition task when the parameters are barely discernable and sample sizes are sufficiently large. Other computational features (e.g., difference

entropy [19] and other higher order features) are eligible but require further investigation.

C. The Neural Network Structure Specifically Designed for the Extracted Boundary Features

1) *Multiple Paths With Circular Networking to Instruct the Neural Network in Analyzing Sector Features:* This paper focuses on neural network design and arrangement of features for effective pattern recognition of ROIs. We designed several neural network connections between the input and the first hidden layers as shown in Fig. 4. In this neural network system, the first layer also functions as a correlation layer that transforms and encodes the signals from input nodes into correlation features for further neural network process. Fig. 4(a)–(c) illustrates the full connection (FC), a self correlation (SC) network, and a neighborhood correlation (NC) network, respectively. Network connections with multiple sectors (i.e., 20° , 30° , 40° , and 50° of the NC) are grouped separately as independent NC paths. In the following study, we used four SC paths for a single sector and thirteen NC paths for four types of multisectors. The method of using the multiple correlation connections was motivated by our research experience in two-dimensional (2-D) convolution neural network (CNN) [(2-D CNN)] where we found that more than ten multiple convolution kernels in the CNN were necessary in the detection of lung nodules and microcalcifications [25].

Compared with 2-D CNN systems, the computation required in the one-dimensional (1-D) CNN (e.g., 144 input features) is relatively small. The combination of the networking paths described earlier for multiple circular path convolution neural network (MCPCCN) was implemented using C programming language. The internal computation algorithm used in the MCPCCN shares the same convolution process as that in the 2-D CNN [25]. Rotation invariance and flip invariance for training the 1-D convolution kernels in the MCPCCN were employed.

The fully connected neural network is a conventional feed-forward MLP neural network. The signals of the fully connected neural network join the other network processes (i.e., SC paths and NC paths) at the single node of the output layer. The signal received at the output node is scaled between zero and one. During the training, zero and one were assigned at the output node to perform backpropagation computation for a nonmass and a mass, respectively. The backpropagation is computed in such a way that the computed incremental errors [see equations (9) and (10)] are retraced into every independent network path. Excluding the output layer, the SC and NC signals are independently arranged and are processed through the 1-D convolution process in the forward propagation. The learning algorithms for all three types of circular network paths are based on the backpropagation training method.

Let $V^0(n', s')$ represents an input signal at the node n' and sector s' . The signal processed through an NC path and to be received at each node, n , on the first hidden layer is

$$N_{j[NC]}^1(n) = \left[\sum_{s'} \sum_{n'} V^0(n', s') \cdot W_{j[NC]}(n', s'; n) \right] + b_{j[NC]}^0(n) \quad (3)$$

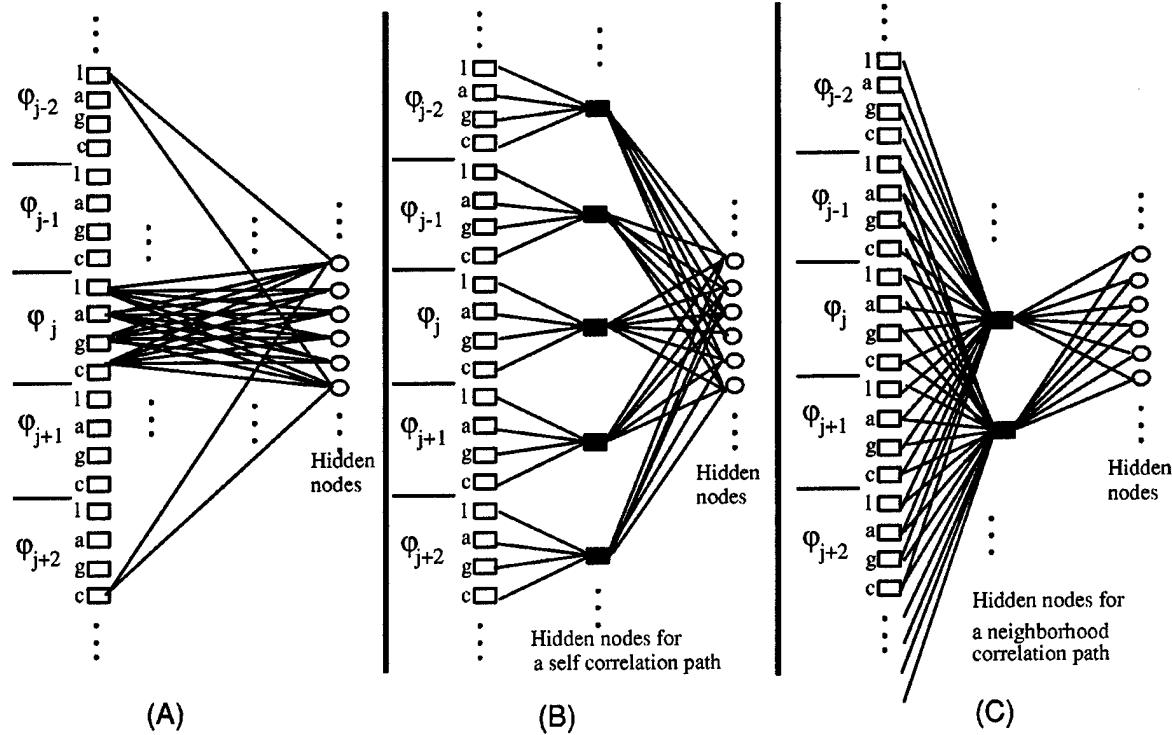


Fig. 4. Three types of network paths connecting the input and the hidden layers in the MCPCCN. (a) FC path. (b) SC path. Each node on the layer connects to a single set of the features (l, a, g, c) for the fan-in and fully connects to the hidden nodes for fan-out. (c) A NC path. Each node on the layer connects to the input nodes of adjacent sectors for the fan-in and fully connects to the hidden nodes for fan-out. The fan-in nets emphasizing SC in (b) and NC in (c) represent convolution weights (i.e., the same type of sectors possess the same set of weighting factors).

where $b_{j[NC]}^0(n)$ represents the bias term and $W_{j[NC]}(n', s'; n)$ is an array associated the 2-D nets that fan-in to a given receiving node, n . Each element of $W_{j[NC]}(n', s'; n)$ is the weight factor connected to node n from node n' sector s' through a NC path, j , and s' covers a range of neighborhood sectors corresponding to each type of NC path. Note that multiplications between the input nodes and connecting weights are computed first followed by taking the sum of the products for those nodes and sectors involved. The operation is repeated by shifting the weights from one set of sectors to the next. The procedure involving array multiplication passing through every sector is referred as the 1-D convolution operation that takes place in the sector dimension. The signal processed through an SC path and to be received at a node, n , on the first hidden layer is a special case of an NC path when s' only covers one sector

$$N_{i[SC]}^1(n) = \left[\sum_{n'} V^0(n', s') \cdot W_{i[SC]}(n'; n) \right] + b_{i[SC]}^0(n) \quad (4)$$

where $W_{i[SC]}(n'; n)$ is the weight factor connected to n from node n' through a SC path, i , regardless of the sectors. A total of 18 paths (1 FC, 4 SC paths, and 13 NC paths for four types of multisectors) were used in our experiment described later. Nevertheless, the signals processed through a path and to be received at each node, n , on the first hidden layer is

$$V_P^1(n) = S(N_P^1(n)) \quad (5)$$

where p is one of the network paths and $S(z)$ is a sigmoid function given by

$$S(z) = \frac{1}{1 + \exp(-z)}. \quad (6)$$

The sigmoid function would produce modulated values ranging from zero to one. The signals on other hidden layers in each path are processed the same as a conventional fully connected neural network. Other than the first hidden layer, the receiving signals at a hidden layer, l , collected from the previous hidden layer, $l-1$, are merged from the nodes in the last layer and are given by

$$V^l(n) = S(N^l(n)) = S \left(\sum_{n'} V^{l-1}(n') \cdot W^{l-1}(n'; n) + b^{l-1}(n) \right) \quad (7)$$

where n' and n denote the nodes at layers $l-1$ and l , respectively.

Let the t th change of the weight be $\Delta W_p^l(n', s'; n)$ and the t th change of the bias be $\Delta b^l(t)$. The error function is defined as

$$E = \frac{1}{2} (T - O)^2 \quad (8)$$

where T and O denote the target output value and the actual output value, respectively when the input values $V^0(n', s')$, are

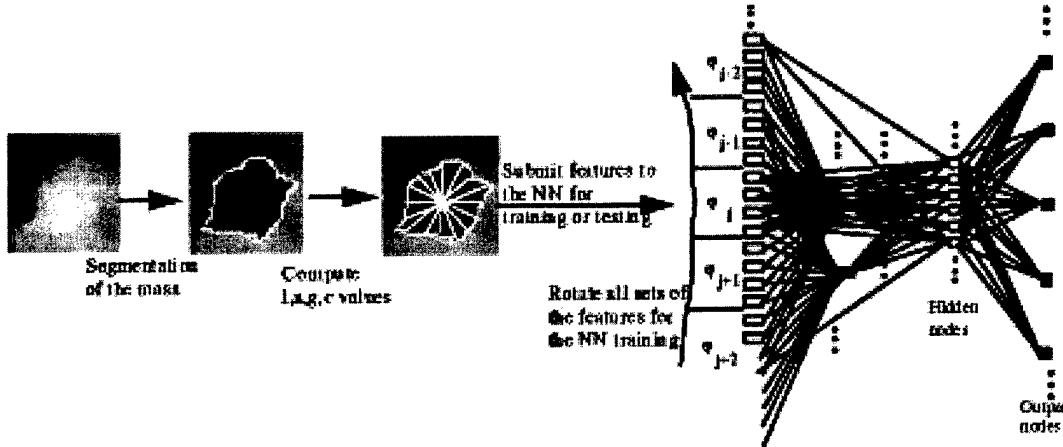


Fig. 5. A schematic diagram, showing the MCPCCN and sector features of masses, that was used in the following study.

entered in the network. In this model, the error backpropagation algorithm, which updates the kernel weights, is given below

$$\Delta W_p^l[t+1] = \eta \left(\sum_n \sum_s \delta_p^{l+1}(n', s'; n, s) \cdot V_p^{l+1}(n, s) + \alpha \Delta W_p^l[t] \right) \quad (9)$$

$$\Delta b_p^l[t+1] = \eta \sum_n \sum_s \delta_p^{l+1}(n', s'; n, s) + \alpha \Delta b_p^l[t] \quad (10)$$

$$\delta_p^l(n', s'; n, s) = S'(N_p^l(n', s')) \left(\sum_n \sum_s \delta_p^{l+1}(n, s) \cdot W_p^{l+1}(n, s) \right). \quad (11)$$

In the case of the last layer

$$\delta^L(n) = S'(N^L(n)) (T(n) - O(n)) \quad (12)$$

where $S'(z)$, η , α , and T denote the derivative of $S(z)$, the learning rate, the weighting factor contributed by the momentum term, and the desired output image, respectively. Furthermore, s or $s' = 1$ and $p = 1$ when $l \neq 0$.

During the training, we added an isotropic constraint to the weights of the 1-D convolution kernels so that

$$W_q^0(n, -s) = W_q^0(n, s) \quad (13)$$

where q is not the fully connected path. These additional constraints are used to induce the kernels functioning as correlation processing filters and could facilitate the algorithm in searching for an appropriate filter.

2) *Resampling the Training Set Through Utilization of Rotation and Flip Invariance of the Features:* In this neural network model, there are no starting and ending sectors. The forward and backpropagation computation can start from any sector. Considering a flipped patch, the characteristics of mass feature should remain the same. To take advantage of this flip invariance, the same numerical target value can be assigned at the output node

for the flipped image patch in order to double the amount of cases during training.

Since we designed a 10° increment for each rotation, every SC or NC path would process through 36 times using the defined features for each image patch. To simplify this network computation, we shifted one small sector (four nodes) on the input layer at a time to conduct the circular convolution process with the SC and NC kernels in the following experiments. By reversing the sequence of the sector, one can train the flipped version of the suspicious masses. Hence, using the properties of the rotation invariance and flip invariance for the neural network training literally increases the number of the training set by a factor of 72.

In summary, we have developed a complete detection procedure for the automatic recognition of mammographic masses including background adjustment, contrast enhancement, ROI segmentation, feature extraction, and MCPCCN system with a training method. Fig. 5 shows a flow diagram for the essential sections of the computational procedures.

IV. EXPERIMENTS AND RESULTS

As described in Section III-A, the 200 mammograms were selected from the MIAS database and the BAMC database for the study. Of the 200 mammograms, 50 mammograms are normal, and each of the 150 abnormal mammograms contains at least one mass case of varying size, subtlety, and location. Both the crano-caudal (CC) and medio-lateral oblique (MLO) projection views were used. The films were digitized with a computer format of $2048 \times 2500 \times 12$ bits (for an $8'' \times 10''$ area where each image pixel represents $100 \mu\text{m}$ square). Ninety-one mammograms, either a CC or an MLO view film, were selected from 91 patient film jackets. No two mammograms were selected from the same patient. All the digitized mammograms were miniaturized to $512 \times 625 \times 12$ bits using 4×4 pixel averaging before the method was applied. According to radiologists, the size of small masses is 3–15 mm in effective diameter. A 3-mm object in an original mammogram occupies 30 pixels in a digitized image with a $100\text{-}\mu\text{m}$ resolution. After reducing the image size by four times, the object will occupy the range of about 7–8 pixels. The object with the size of seven pixels

is expected to be detectable by any computer algorithm. After preprocessing and an object screening based on the circularity test and the size test (between 3 and 30 mm), a total of 125 suspicious areas were selected from the testing mammograms (91 cases) for this study. Specifically, the screening procedure of reducing FPs involves two steps: 1) image patches with circularity less than 0.25 or diameter greater than 30 mm were eliminated and 2) using probability modular neural network to rule out the majority of FPs. Of the 125 suspicious areas, 75 ROIs contained masses based on corresponding biopsy reports with one experienced radiologist reading. Of 75 masses, 39 were malignant and 36 were benign. This set of ROIs was used in [19] and discussed in [19, Fig. 6 and Table II].

A. Experiment 1

Of the 125 suspicious areas, we randomly selected 54 computer-segmented ROIs where 30 patches were matched with the radiologist's mass identification and 24 were not. This database was used to train two neural network systems: 1) a conventional three-layer neural network and 2) the proposed MCPCNN training method using the same neural network learning algorithm. The structure of the MCPCNN was described earlier. In the study, we used one fully connected path, four SC paths, four NC paths covering two sectors, four NC paths covering three sectors, three NC paths covering four sectors, and two NC paths covering five sectors in the first step network connection for the MCPCNN. All paths in the neural network have their hidden layers. Only one hidden layer per path was used. Both neural network systems were trained by the error backpropagation algorithm by feeding the features from the input layer and registering the corresponding target value at the output node. Completion of the training was determined by the mean square error [i.e., $\sum_{i=1}^N (T_i - O_i)^2 / N$, where N is number of samples] when it was approximately reduced to 3×10^{-5} . Once the training of the neural networks was completed, we then used the remaining 71 computer segmented ROIs for the testing. Forty-five out of 71 ROIs were masses and 26 ROIs were not. Neither the images nor their corresponding patients in the testing set could be found in the training set. The neural network output values were fed into the LABROC4 program [29] for the performance evaluation. The results indicated that the areas (A_z) under the receiver operating characteristic (ROC) curves were 0.7869 ± 0.0536 and 0.8443 ± 0.0457 using the conventional neural network (MLP) and the MCPCNN, respectively. The ROC curves of these two neural network systems are shown in Fig. 6(a). The A_z value was 0.7869 ± 0.0536 when using the MLP method with 125 hidden nodes. The performance of the MLP remains about the same at 0.7809 ± 0.0551 of A_z using the same neural network parameters but with 30 hidden nodes.

We also invited another senior mammographer to conduct an observer study using the ROC study protocol. The mammographer was asked to rate each patch using a numerical scale ranging from zero to ten for its likelihood of being a breast mass. The image patches were displayed on a SUN monitor (Model: GDM-20D10). The image size shown on the monitor was reduced to approximately $7'' \times 9''$ as compared with the original film size ($8'' \times 10''$). These 71 numbers were also fed into the LABROC4 program. The A_z of the mammographer's perfor-

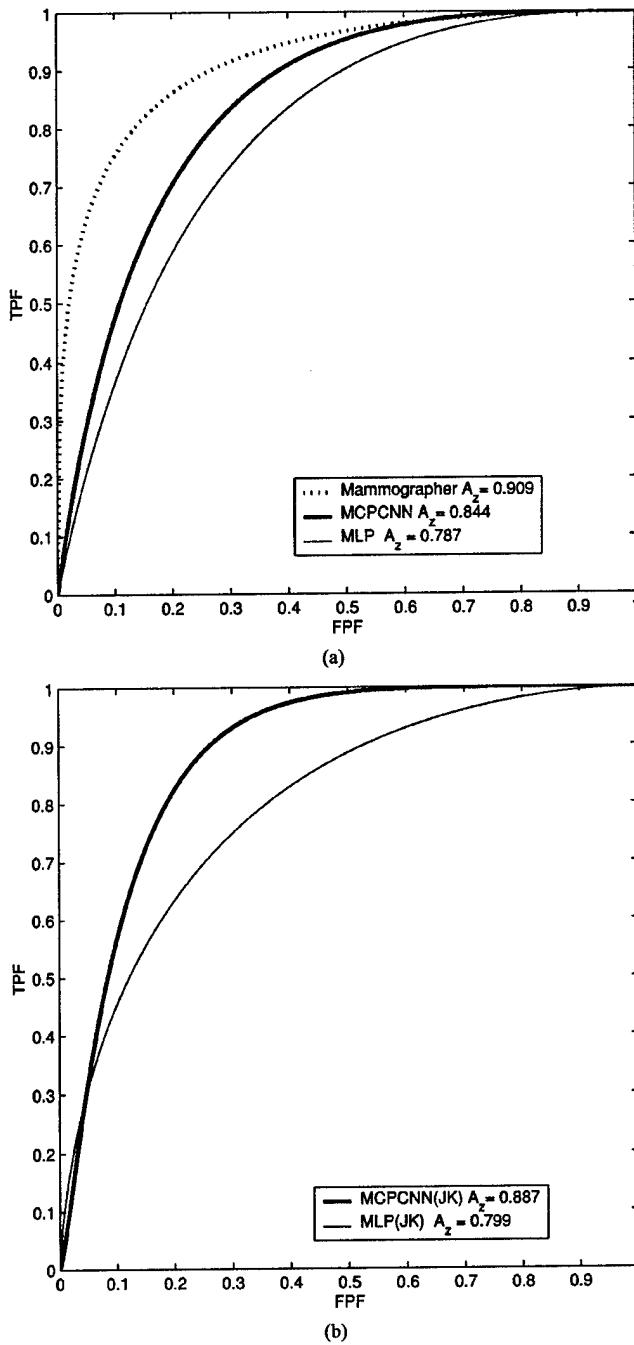


Fig. 6. The ROC curves obtained from corresponding experiments. (a) Shows that the performance of MCPCNN training method is superior to that of the conventional MLP method. The highest curve is the ROC performance of the senior mammographer. (b) Shows that the ROC results were increased using the leave-one-case-out procedure in both neural network systems. The MCPCNN still showed higher performance than conventional MLP method.

mance on this set of test cases was 0.909 ± 0.0340 . The corresponding ROC curve is also shown in Fig. 6(a).

B. Experiment 2

We also conducted a leave-one-case-out experiment (i.e., jackknife procedure) using the same database. In this experiment, we used those image patches extracted from 90

TABLE I
ROC PERFORMANCE OF THE TEST METHODS IN DISTINGUISHING TRUE AND FALSE MASSES

	Comparative Analyses of Methods	A _z of Method (1)	A _z of Method (2)	P Values	Statistical Significance
Experiment 1	(1) Radiologist vs. (2) MCPCNN	0.909 ± 0.0340	0.8443 ± 0.0457	0.1855	No
	(1) Radiologist vs. (2) MLP	0.909 ± 0.0340	0.7869 ± 0.0536	0.0447	Yes
	(1) MCPCNN vs. (2) MLP	0.8443 ± 0.0457	0.7869 ± 0.0536	0.1344	No
Experiment 2	(1) MCPCNN vs. (2) MLP	0.8866 ± 0.0289	0.7985 ± 0.0394	0.0241	Yes

mammograms (one mammogram per case) for the training and used the image patches (most of them are single) extracted from the remaining one mammogram as test objects. The procedure was repeated 91 times to allow every ROI extracted from each mammogram to be tested in the experiment. For each individual ROI, the computed features were identical to those used in Experiment 1. Again, the training was stopped when the mean square error value approximately equal to 3×10^{-5} . Both neural network systems were independently trained and evaluated with the same procedure. The results indicated that the A_z values were 0.7985 ± 0.0394 and 0.8866 ± 0.0289 using the conventional neural network (MLP) and the MCPCNN, respectively. The performance of the MLP decreased to an A_z of 0.7608 ± 0.0429 using the same neural network parameters but with 30 hidden nodes. Fig. 6(b) shows the ROC curves of these two neural network systems using the leave-one-case-out procedure [30] in the experiment.

We also used CLABROC program [31] to analyze the ROC data and compare the ROC results. The results and their statistical significances using two tailed p value of 0.05 as the threshold are shown in Table I. The radiologist's performance is greater than conventional neural network system with a p value of 0.0447 in the first experiment. The MCPCNN was also proven to be superior to the MLP with a statistically significant result ($p = 0.0241$).

V. DISCUSSION

It is known in the field of artificial intelligence that the key factors in pattern recognition are: 1) effective methods in the extraction of features and 2) classification methods for the extracted features. In this study, we showed that the training method designed to guide the analyzer is also an important factor for a pattern recognition task. Though this finding is not new, the research of developing training methods for various pattern recognition tasks has not been established in the field of medical imaging. Our studies demonstrated that with proper network connections and task-oriented guidance, organized features would assist the neural network in performing the task.

Technically speaking, a feed-forward MLP neural network provides an integrated process for classification and sometimes for feature extraction. The output values of the hidden nodes can be interpreted as a reorganized set of features presented to the output layer for classification. The drawback of the MLP is, the user has a very little control and little understanding about the network learning. The MCPCNN is a network design that partially remedies these issues and is applicable for any pattern recognition task associated with ROIs. The MCPCNN (a

member of the CNN family) possesses shared weights in the hidden layer(s) that act as filter kernels for extracting correlated features. With a higher resolution mammogram, a finer sector ($<10^\circ$) would be preferred for the analysis mass, especially for the study of classification of masses. During forward and back-propagation training, the kernels would comply with both signals from input and output layers for all training cases, so as to maximize the classification performance. We do not recommend using 2D CNN for the detection of masses because the mass sizes vary from a few millimeters to 4 cm or even larger. It would require a large fixed size to cover the maximum mass size when using the 2-D CNN. The varieties of mass shapes and potential long spiculated patterns make the use of the 2-D CNN not practical. Since the MCPCNN processes the features computed from sectors, it does not limit the sizes of its ROIs. Best of all, the MCPCNN also has the ability to classify partially obscured masses. The 2-D CNN, however, would be more appropriate for the detection of microcalcifications and small lung nodules.

As far as the research in the detection of masses is concerned, we have shown that use of MCPCNN with sector features is an effective approach. Since the MCPCNN coordinates the input data and performs correlation between features of adjacent sectors in the first stage of data processing, the internal neural network learning algorithm can be changed if a learning algorithm is found to be more effective. In fact, the MCPCNN is a technique that can effectively classify features arranged in the polar coordinate system. A technique using the rubber band straightening transformation, independently developed by Sahnier *et al.* [11], for the detection of masses also employs a similar concept in extracting feature and/or texture in the polar coordinate system. We believe that integration of features and texture values computed at small sectors will be the research trend in mass detection and tumor classification.

VI. CONCLUSION

In the clinical course of detecting masses, mammographers usually evaluate the surrounding background of a radiodense area when an ROI is suspected. In this study, we simulated this fundamental concept with a neural network system (i.e., MCPCNN). In order for the MCPCNN to function, boundary features of the suspicious region in each radial sector were computed. We found that the MCPCNN is capable of analyzing correlated features within the sector and between adjacent sectors, which led to an improvement in detecting mammographic masses.

Through this study, we found that the selected features are somewhat effective in the detection of masses. These features

were "computationally translated" from the qualitative descriptors of BI—RAD. These features can be extended for the improvement of the mass detection, but this task is beyond the scope of this paper. With the preliminary studies shown above, we found the MCPCNN coupling with the proposed training method produced greater results than the conventional neural network. We found that the performances of both neural network systems were improved in Experiment 2. This may have occurred due to the number of training samples that was increased from 54 to 124. In Experiment 2, the A_z value was improved by 0.042 using the MCPCNN, which was higher than the A_z difference of 0.012 obtained by the conventional training method. The results implied that the MCPCNN learned more effectively than the conventional neural network when the number of training cases was increased. With the use of a larger database and advanced texture features proposed by others, it is expected that the performance of MCPCNN should be significantly improved. This paper does not intend to claim the best mass detection system, in comparison to similar systems; but rather its goal is to report a potentially better neural network structure for analyzing a set of mass features.

ACKNOWLEDGMENT

A part of the database, used in the study, was provided by Dr. R. Shah of Brooke Army Medical Center. The LABROC4 and CLABROC programs were written by Dr. C. E. Metz and his colleagues at the University of Chicago.

REFERENCES

- [1] L. Nystrom, L. E. Rutqvist, S. Wall, A. Lindgren, M. Lindqvist, and S. Ryden *et al.*, "Breast cancer screening with mammography: Overview of Swedish randomized trials," *Lancet*, vol. 341, pp. 973–978, 1993.
- [2] S. Shapiro, "Screening-assessment of current studies," *Cancer*, vol. 74, pp. 231–238, 1994.
- [3] L. Tabar, G. Fagerberg, S. Duffy, N. E. Day, A. Gad, and O. Grontoft, "Update of the Swedish two-country program of mammographic screening for breast cancer," *Radiol. Clin. N. Amer.: Breast Imag.—Current Status Future Directions*, vol. 30, pp. 187–210, 1992.
- [4] D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms," *IEEE Trans. Med. Imag.*, vol. 9, p. 233, Sept. 1990.
- [5] R. Zwijsgeelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. M. Boggis, "Model-based detection of spiculated lesions in mammograms," *Med. Image Anal.*, vol. 3, no. 1, pp. 39–62, 1999.
- [6] N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," *Med. Phys.*, vol. 23, no. 10, pp. 1685–1696, 1996.
- [7] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodlett, "Classification of mass and normal breast tissues: A convolution neural network classifier with spatial domain and texture images," *IEEE Trans. Med. Imag.*, vol. 15, pp. 598–610, Oct. 1996.
- [8] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodlett, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Med. Phys.*, vol. 25, no. 4, pp. 516–526, 1998.
- [9] L. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Trans. Med. Imag.*, vol. 18, pp. 1178–1187, Dec. 1999.
- [10] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized detection of malignant tumors on digital mammograms," *IEEE Trans. Med. Imag.*, vol. 18, pp. 369–378, May 1999.
- [11] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodlett, "Computerized characterization of masses on mammograms: The rubber band straightening transform and textures analysis," *Med. Phys.*, vol. 25, no. 4, pp. 516–526, 1998.
- [12] M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Trans. Med. Imag.*, vol. 17, pp. 510–517, Aug. 1998.
- [13] D. D. Adler, "Breast Masses: Differential Diagnosis," in *ARRS Categorical Course Syllabus on Breast Imaging*, S. A. Feig, Ed. Reston, VA: Amer. Roent. Ray Soc., 1988, p. 31.
- [14] M. J. Homer, "Imaging features and management of characteristically benign and probably benign breast lesions," *Radiol. Clin. N. Amer.*, vol. 25, p. 939, 1987.
- [15] S. Pohlman, K. A. Powell, N. A. Obuchowski, W. A. Chilcote, and S. Grundfest-Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," *Med. Phys.*, vol. 23, no. 8, pp. 1337–1345, 1996.
- [16] M. Moskowitz, "Circumscribed lesions of the breast," in *Diagnostic Categorical Course in Breast Imaging*, M. Moskowitz, Ed. Oak Brook, IL: Radiol. Soc. N. Amer., 1986, p. 31.
- [17] E. A. Sickles, "The rule of multiplicity and the developing density sign," in *ARRS Categorical Course Syllabus on Breast Imaging*, S. A. Feig, Ed. Reston, VA: Amer. Roent. Ray Soc., 1988, p. 177.
- [18] H. Li, Y. Wang, K.-J. R. Liu, S.-C. B. Lo, and M. T. Freedman, "Computerized radiographic mass detection—Part I: Lesion site selection by morphological enhancement and contextual segmentation," *IEEE Trans. Med. Imag.*, pp. 289–301, Apr. 2001.
- [19] ———, "Computerized radiographic mass detection—Part II: Decision support by featured database visualization and modular neural networks," *IEEE Trans. Med. Imag.*, pp. 302–313, Apr. 2001.
- [20] *Breast Imaging—Reporting and Data System*. Reston, VA: Ame. Coll. Radiol., 1993.
- [21] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cernez, S. Kok, P. Taylor, D. Betal, and J. Savage, "The mammographic images analysis society digital mammogram database," in *Excerpta Medica*, ser. Int. Congr., 1994, vol. 1069, (e-mail for inquiry: mias@sv1.smb.man.ac.uk), pp. 375–378.
- [22] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall., 1999.
- [23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representation by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: M.I.T. Press, 1986, vol. 1, Foundation, ch. 8, pp. 318–362.
- [24] S.-C. B. Lo, S. L. Lou, J. S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications to lung nodule detection," *IEEE Trans. Med. Imag.*, vol. 14, pp. 711–718, Dec. 1995.
- [25] S.-C. B. Lo, H. P. Chan, J. S. Lin, H. Li, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network for medical image pattern recognition," *Neural Networks*, vol. 8, no. 7/8, pp. 1201–1214, 1995.
- [26] H. P. Chan, S.-C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided diagnosis of mammographic microcalcifications: Pattern recognition with an artificial neural network," *Med. Phys.*, vol. 24, no. 10, pp. 1555–1567, 1995.
- [27] Y. Wu, K. Doi, M. L. Giger, and R. M. Nishikawa, "Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks," *Med. Phys.*, vol. 19, pp. 555–560, 1992.
- [28] Y. Wu, M. T. Freedman, S.-C. B. Lo, R. A. Zuurbier, A. Hasegawa, and S. K. Mun, "Classification of microcalcifications in radiographs of pathological specimen for the diagnosis of breast cancer," *Acad. Radiol.*, vol. 2, pp. 199–204, 1995.
- [29] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statist. Med.*, vol. 17, pp. 1033–1053, 1998.
- [30] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Trans. Pattern Anal. Machine Intell.*, pp. 873–885, Aug. 1989.
- [31] C. E. Metz, P.-L. Wang, and H. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging*, F. Deconinck, Ed. The Hague, The Netherlands: Martinus Nijhoff, 1984, vol. PAMI-II, pp. 432–445.

Information-Theoretic Matching of Two Point Sets

Yue Wang, Kelvin Woods, and Maxine McClain

Abstract—This paper describes the theoretic roadmap of least relative entropy matching of two point sets. The novel feature is to align two point sets without needing to establish explicit point correspondences. The recovery of transformational geometry is achieved using a mixture of principal axes registrations, whose parameters are estimated by minimizing the relative entropy between the two point distributions and using the expectation-maximization algorithm. We give evidence of the optimality of the method and we then evaluate the algorithm's performance in both rigid and nonrigid image registration cases.

Index Terms—Finite normal mixture, image registration, information theory, neural computation.

I. INTRODUCTION

THE ESTIMATION of transformational geometry from two point sets is an essential step to medical imaging and computer vision [1], [2]. The task is to recover a matrix representation requiring a set of correspondence matches between features in the two coordinate system [3]. Assume two point sets $\{\mathbf{p}_{iA}\}$ and $\{\mathbf{p}_{iB}\}; i = 1, 2, \dots, N$ are related by

$$\mathbf{p}_{iB} = \mathbf{R}\mathbf{p}_{iA} + \mathbf{T} + \mathbf{N}_i \quad (1)$$

where \mathbf{R} is a rotation matrix, \mathbf{T} is a translation vector, and \mathbf{N}_i is a noise vector. Given $\{\mathbf{p}_{iA}\}$ and $\{\mathbf{p}_{iB}\}$, Arun *et al.* present an algorithm for finding the least-squares solution of \mathbf{R} and \mathbf{T} , which is based on the decoupling of translation and rotation and the singular value decomposition of a 3×3 cross-covariance matrix [3].

The major limitation of the present method is twofold: 1) while feature matching methods can give quite accurate solutions, obtaining correct correspondences of features is a hard problem, especially in the cases of images acquired using different modalities or taken over a period of time and 2) a rigidity assumption is heuristically imposed, leading to the incapability of handling situations with nonrigid deformations. One popular method that does not require correspondences is the principal axes registration (PAR) [1], which is based on the relatively stable geometric properties of image features, i.e., the

Manuscript received October 20, 2000; revised April 20, 2002. This work was supported in part by the Department of Defense under Grant DAMD17-98-8045 and the National Institutes of Health under Grant R21RR12784. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Brian L. Evans.

Y. Wang is with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA and also with the Department of Radiology and Radiological Science, The Johns Hopkins University School of Medicine, Baltimore, MD 21205 USA (e-mail: wang@pluto.ee.cua.edu).

K. Woods and M. McClain are with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA.

Publisher Item Identifier 10.1109/TIP.2002.801120.

geometric information contained in these stable image features is often sufficient to determine the transformation between images [2].

In this paper, we first discuss the optimality of PAR in a maximum likelihood (ML) sense. The novel feature is to align two point sets without needing to establish explicit point correspondences. We then propose a somewhat different approach for recovering transformational geometry of nonrigid deformations. That is, rather than using a single transformation matrix which gives rise to a large registration error, we attempt to use a mixture of principal axes registrations (mPAR), whose parameters are estimated by minimizing the relative entropy between the two point distributions and using the expectation-maximization algorithm. We demonstrate the principle of the method for both rigid and nonrigid image registration cases.

II. THEORY AND METHOD

A. Optimality of PAR

As suggested by information theory [4], we note that the control point sets in two images can be considered as two separate realizations of the same random source. Therefore, we do not need to establish point correspondences to extract the transformation matrix. In other words, if we denote by $P_{\{\mathbf{p}_i\}}$ the distribution of the control point set in an image, we have the simple relationship

$$P_{\{\mathbf{p}_{jB}\}} = P_{\{\mathbf{R}\mathbf{p}_{iA} + \mathbf{T}\}} + \nu \quad (2)$$

where ν is the noise component (caused by misalignment) [2]. The probability distributions can be computed independently on each image without any need to establish feature correspondences, and given the two distributions of the control point sets in the two images, we can recover the transformation matrix in a simple fashion [2], as we now describe.

From observation of the distributions, we can estimate \mathbf{R} and \mathbf{T} by minimizing the relative entropy (Kullback-Leibler distance) between $P_{\{\mathbf{p}_{jB}\}}$ and $P_{\{\mathbf{R}\mathbf{p}_{iA} + \mathbf{T}\}}$, i.e.,

$$\arg \min_{\mathbf{R}, \mathbf{T}} D(P_{\{\mathbf{p}_{jB}\}} \| P_{\{\mathbf{R}\mathbf{p}_{iA} + \mathbf{T}\}}) \quad (3)$$

where D denotes the relative entropy measure. We have previously shown the relationship between the negative log joint likelihood and the relative entropy as (Theorem 1) [5]

$$\begin{aligned} -\frac{1}{N_B} \log \mathcal{L}(P_{\{\mathbf{R}\mathbf{p}_{iA} + \mathbf{T}\}}(\mathbf{p}_{jB})) \\ = H(P_{\{\mathbf{p}_{jB}\}}) + D(P_{\{\mathbf{p}_{jB}\}} \| P_{\{\mathbf{R}\mathbf{p}_{iA} + \mathbf{T}\}}) \end{aligned} \quad (4)$$

where H denotes the entropy measure. Thus, minimizing $D(P_{\{\mathbf{p}_{jB}\}} \| P_{\{\mathbf{R}\mathbf{p}_{iA} + \mathbf{T}\}})$ is equivalent to maximizing $\log \mathcal{L}(P_{\{\mathbf{R}\mathbf{p}_{iA} + \mathbf{T}\}}(\mathbf{p}_{jB}))$. Following the same strategy to

decouple translation and rotation as in [3], we can define a new data point by $\mathbf{q}_{iA} = \mathbf{p}_{iA} - \mathbf{p}_A^0$ and $\mathbf{q}_{jB} = \mathbf{p}_{jB} - \mathbf{p}_B^0$, where \mathbf{p}_A^0 and \mathbf{p}_B^0 are the centroids of $\{\mathbf{p}_{iA}\}$ and $\{\mathbf{p}_{jB}\}$, respectively. Then the ML estimator of \mathbf{R} is defined by

$$\arg \max_{\mathbf{R}} \log \mathcal{L}(P_{\{\mathbf{q}_{iA}\}}(\mathbf{q}_{jB})) \quad (5)$$

and $\mathbf{T} = \mathbf{p}_B^0 - \mathbf{R}\mathbf{p}_A^0$.

In the case of principal axes technique, we assume a Gaussian model for $P_{\{\mathbf{q}_{iA}\}}$ and $P_{\{\mathbf{q}_{jB}\}}$. Therefore,

$$\begin{aligned} \log \mathcal{L} & \left(\frac{N_A^{1/2}}{(2\pi)^{3/2} |\mathbf{RC}_A \mathbf{R}^t|^{1/2}} \right. \\ & \times \exp \left(-\frac{1}{2} \mathbf{q}_{jB}^t (\mathbf{RC}_A \mathbf{R}^t)^{-1} \mathbf{q}_{jB} \right) \Big) \\ & = \log \frac{N_A^{1/2} N_B}{(2\pi)^{3/2}} - \frac{1}{2} \log |\mathbf{RC}_A \mathbf{R}^t| \\ & - \frac{1}{2N_B} \sum_{j=1}^{N_B} \mathbf{q}_{jB}^t (\mathbf{RC}_A \mathbf{R}^t)^{-1} \mathbf{q}_{jB} \end{aligned} \quad (6)$$

where the superscript t denotes matrix transposition, $\mathbf{C}_()$ denotes the auto-covariance matrix

$$\mathbf{C}_A \triangleq \frac{1}{N_A} \sum_{i=1}^{N_A} \mathbf{q}_{iA} \mathbf{q}_{iA}^t \text{ or } \mathbf{C}_B \triangleq \frac{1}{N_B} \sum_{j=1}^{N_B} \mathbf{q}_{jB} \mathbf{q}_{jB}^t \quad (7)$$

and N_A and N_B are the sizes of the point sets $\{\mathbf{q}_{iA}\}$ and $\{\mathbf{q}_{jB}\}$ respectively. By taking the derivative of (6) with respect to \mathbf{R} and setting it equal to zero, we have the ML equation (see hints in Appendix) [6]

$$\mathbf{C}_B = \mathbf{RC}_A \mathbf{R}^t. \quad (8)$$

Now let the eigenvalue decompositions of \mathbf{C}_A and \mathbf{C}_B be

$$\mathbf{C}_A = \mathbf{U}_A \Lambda_A \mathbf{U}_A^t, \quad \mathbf{C}_B = \mathbf{U}_B \Lambda_B \mathbf{U}_B^t \quad (9)$$

where \mathbf{U}_A and \mathbf{U}_B are 3×3 orthonormal matrices and Λ_A and Λ_B are 3×3 diagonal matrices with nonnegative elements. Note that the transformation \mathbf{U} consists of the orthonormal set of eigenvectors of \mathbf{C} , and matrix Λ contains eigenvalues λ_m of \mathbf{C} for $m = 1, 2, 3$. Then, we assign

$$\mathbf{R} = \mathbf{U}_B \mathbf{K} \mathbf{U}_A^t \quad (10)$$

where \mathbf{K} is a 3×3 diagonal matrix with element $k_m = \sqrt{\lambda_{mB}/\lambda_{mA}}$, the right side of ML (8) becomes

$$\mathbf{RC}_A \mathbf{R}^t = \mathbf{U}_B \mathbf{K} \mathbf{U}_A^t \mathbf{U}_A \Lambda_A \mathbf{U}_A^t \mathbf{U}_A \mathbf{K} \mathbf{U}_B^t = \mathbf{U}_B \Lambda_B \mathbf{U}_B^t$$

which equals exactly the left side of ML (8). Thus, among all 3×3 orthonormal matrices, \mathbf{R} defined by (10) that also includes a scaling matrix \mathbf{K} [1], maximizes the joint log likelihood in (6). So far, we have verified the optimality of PAR techniques.

B. Formulation of mPAR

However, because of its global linearity, the application of PAR is necessarily somewhat limited. An alternative paradigm

is to model a multimodal control point set with a collection of local linear models [7]. The method is a two-stage procedure: a soft partitioning of the data set followed by estimation of the principal axes within each partition [8]. Recently there has been considerable success in using standard finite normal mixture (SFNM) to model the distribution of a multimodal data set [5], and the association of a SFNM distribution with PAR offers the possibility of being able to register two images through a mixture of probabilistic principal axes transformations [8].

Assume that there are K_0 control point clusters, where each control point cluster defines a transformation $\{\mathbf{R}_k, \mathbf{T}_k\}$. Thus for a pixel \mathbf{p}_{nA} , its new locations, corresponding to each of the transformations, are $\mathbf{p}_{nk} = \mathbf{R}_k \mathbf{p}_{nA} + \mathbf{T}_k$ for $k = 1, \dots, K_0$. Further assume that the control point set defines a SFNM distribution

$$f(\mathbf{p}_i) = \sum_{k=1}^{K_0} \alpha_k g(\mathbf{p}_i | \boldsymbol{\mu}_k, \mathbf{C}_k) \quad (11)$$

where g is the Gaussian kernel with mean vector $\boldsymbol{\mu}_k$ and auto-covariance matrix \mathbf{C}_k , and α_k is the mixing factor which is proportional to the number of control points in cluster k . For each of the control point sets $\{\mathbf{p}_{iA}\}$ and $\{\mathbf{p}_{iB}\}$, the mixture is fit using the expectation-maximization (EM) algorithm [5]. The E step involves assigning to the linear models contributions from the control points; the M step involves re-estimating the parameters of the linear models in the light of this assignment [8].

E-Step

$$z_{ik}^{(l)} = \frac{\alpha_k^{(l)} g(\mathbf{p}_i | \boldsymbol{\mu}_k^{(l)}, \mathbf{C}_k^{(l)})}{\int f(\mathbf{p}_i | \boldsymbol{\mu}_k^{(l)}, \mathbf{C}_k^{(l)})}. \quad (12)$$

M-Step

$$\alpha_k^{(l+1)} = \frac{1}{N} \sum_{i=1}^N z_{ik}^{(l)} \quad (13)$$

$$\boldsymbol{\mu}_k^{(l+1)} = \frac{\sum_{i=1}^N z_{ik}^{(l)} \mathbf{p}_i}{\sum_{i=1}^N z_{ik}^{(l)}} \quad (14)$$

$$\mathbf{C}_k^{(l+1)} = \frac{\sum_{i=1}^N z_{ik}^{(l)} (\mathbf{p}_i - \boldsymbol{\mu}_k^{(l)}) (\mathbf{p}_i - \boldsymbol{\mu}_k^{(l)})^t}{\sum_{i=1}^N z_{ik}^{(l)}}. \quad (15)$$

For each complete cycle of the algorithm, we first use the “old” set of parameter values to determine the posterior probabilities $z_{ik}^{(l)}$ using (12). These posterior probabilities are then used to obtain “new” values $\alpha_k^{(l+1)}$, $\boldsymbol{\mu}_k^{(l+1)}$, $\mathbf{C}_k^{(l+1)}$ and using (13)–(15). The algorithm cycles back and forth until the value of relative entropy between the data histogram and mixture model $D(P_{\{\mathbf{p}_i\}} \| f(\mathbf{p}_i))$ reaches its saturation point, for $\{\mathbf{p}_{iA}\}$ and $\{\mathbf{p}_{iB}\}$, respectively. Our experience indicates that 20 iterations should be sufficient to reach such point, although the number of iterations may vary from case to case occasionally.

Thus the statistical membership of pixel \mathbf{p}_{nA} belonging to each of the control (point) clusters can be derived by

$$z_{nk} = P(\mathbf{R}_k, \mathbf{T}_k | \mathbf{p}_{nA}) = \frac{\alpha_k g(\mathbf{p}_{nA} | \boldsymbol{\mu}_k, \mathbf{C}_k)}{\int f(\mathbf{p}_{nA})} \quad (16)$$

i.e., the posterior probability of $\{\mathbf{R}_k, \mathbf{T}_k\}$ given \mathbf{p}_{nA} . We can define the mPAR transformation as

$$\begin{aligned} \mathbf{p}_n &= \sum_{k=1}^{K_0} z_{nk} \mathbf{p}_{nk} \\ &= \sum_{k=1}^{K_0} \frac{\alpha_{kA} g(\mathbf{p}_{nA} | \mu_{kA}, \mathbf{C}_{kA})}{f(\mathbf{p}_{nA})} (\mathbf{R}_k \mathbf{p}_{nA} + \mathbf{T}_k) \quad (17) \end{aligned}$$

where $\{\mathbf{R}_k, \mathbf{T}_k\}$ is determined based on $\{(\mu_{kA}, \mathbf{C}_{kA}), (\mu_{kB}, \mathbf{C}_{kB})\}$ that we have estimated in the previous step using the EM algorithm. Note that now we do need the correspondences between the two control (point) clusters for each k . These correspondences may be found, after a global PAR is initially performed, by using a site model approach or a dual-step EM algorithm to unify the tasks of estimating transformation geometry and identifying cluster-correspondence matches [10]. This philosophy for recovering transformational geometry of the nonrigid deformations is similar in spirit to the modular networks in neural computation [5], [7], under which the relative entropy between the two point sets reaches its minimum

$$\arg \min_{\mathbf{R}_k, \mathbf{T}_k} D \left(P_{\{\mathbf{p}_{kB}\}} \| P_{\left\{ \sum_{k=1}^{K_0} z_{ik} (\mathbf{R}_k \mathbf{p}_{nA} + \mathbf{T}_k) \right\}} \right) \quad (18)$$

both globally and locally.

III. RESULT AND DISCUSSION

We first illustrate the application of PAR to the coregistration of human brain scans by magnetic resonance (MR) imaging and positron emission tomography (PET). The purpose of this experiment is to demonstrate that under a rigidity assumption and without knowing control point correspondences, PAR provides a satisfactory solution to multimodality image co-registration. MR scan was performed using a GE Signa 1.5 Tesla system, with 1.5 mm effective slice thickness, zero gap, 124 slices of in-plane 192×256 matrix, and 24 cm field-of-view, to cover vertex to foramen magnum [9]. A thermoplastic mask was prepared to fit the patient's face, and facilitate positioning and repositioning in PET serving as the ground truth. PET images were then obtained using a GE 4096 whole body scanner with 15 slices at the center of the field-of-view. The slice thickness is 6.5 mm and the spatial resolution is 6–7 mm. A typical pair of MR and PET brain images is shown in Fig. 1.

Since PET images are often very noisy (i.e., with high speckle noise), an effective pre-processing is performed, which jointly uses image segmentation and morphological filtering [9], to eliminate background noise and extract the geometric contour of the brain tissue area. For MR images, the skull and scalp do not contribute to the functional activity shown in PET images, we edit MR images to delineate the skull and scalp and successfully separate the brain tissues out from MR head scans. The extracted contours have good edge correspondence to both PET and MR images (see Fig. 1) [9].

The results of PAR show that the angle of rotation relative to the principal axis is -3.90° and -4.91° for PET and MR images respectively. Therefore, the relative angle of rotation of

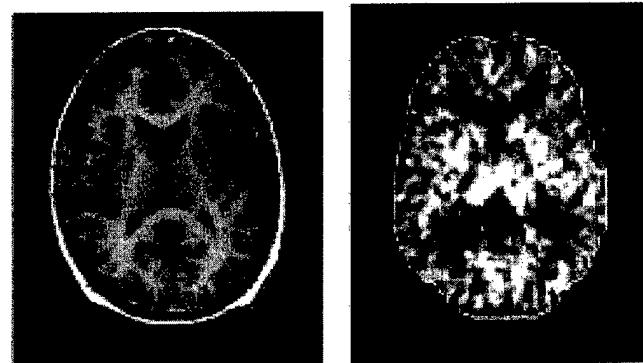


Fig. 1. (Left) MR and (right) PET brain scans, where the extracted contour of brain tissue area from the MR image is overlaid on the registered PET image.

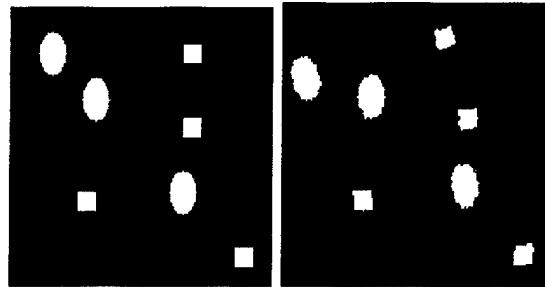


Fig. 2. (Left) Reference phantom image and (right) warped floating phantom image using mPAR method.

MR with respect to PET is -1.01° which agrees closely to our ground truth of -1.2° . In addition, the scaling factor is found to be 2.20, which also matches closely with our ground truth of 2.21. To demonstrate such co-registration, the re-sampled contour of MR brain tissue area is overlaid on the registered PET image as displayed in Fig. 1 (right). In this experiment, the capable nature of the PAR for rigid alignment is evident as the two contours match each other very well.

To evaluate the effectiveness of the mPAR method, we first considered a 150×150 phantom study containing three control objects and four noncontrol objects as seen in Fig. 2. The control objects are ellipses while the noncontrol objects are squares. Each of the control and noncontrol objects are rotated and translated by different amounts. This simulates a nonlinear deformation (nonrigid) between image sets. Three configurations of rotation angles are considered. These configurations are chosen randomly (within certain range) to show the robustness of the proposed algorithm. In each configuration the images are aligned using one, two, or three transformations.

The performance is measured in mean square error (MSE) between the reference and warped images. Our experiments show that registration by one transform on average reduces the MSE by 50% and further reduces another 10% with one additional transform. With a mixture of all the three local transformations, a significant improvement in MSE is achieved with a reduction of approximately 75%. Fig. 2 shows an example of the reference (left) and warped image (right) using all three transformations. The result shows the benefit of using multiple transformations where possible.

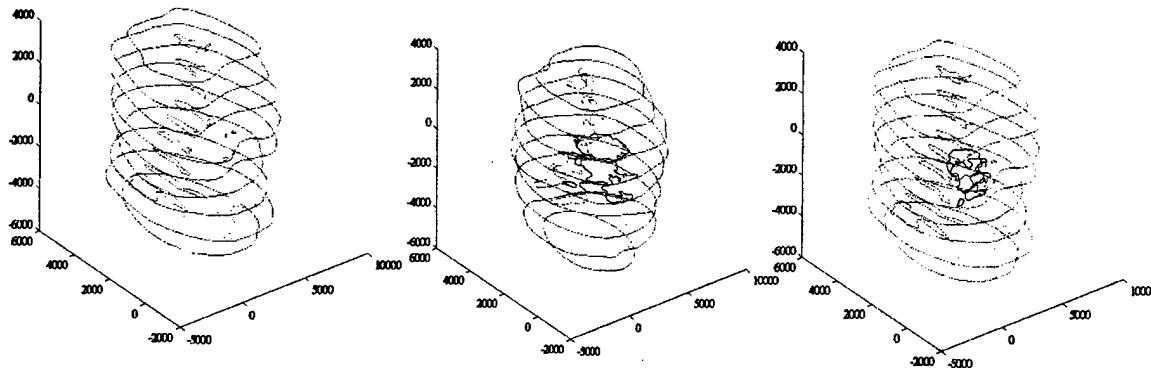


Fig. 3. Result of applying mPAR method to register two 3-D prostate models reconstructed from two different real surgical specimens. Left: reference model. Middle: floating model. Right: the tumor in the floating model has been mapped into the reference model.

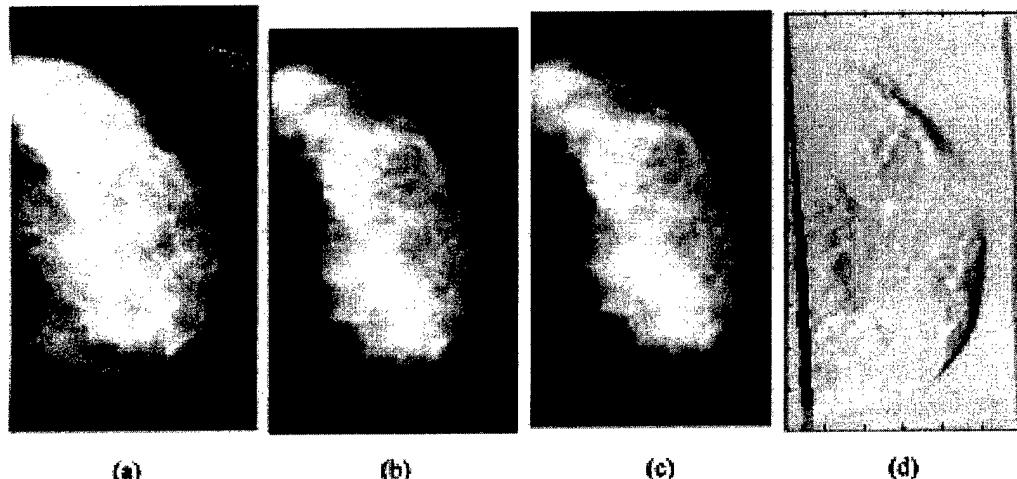


Fig. 4. (a) and (b) Pair of real mammograms used to detect changes of breast structure over time. The mPAR method is used to perform an initial registration based on all the extracted control points in the two images. Based on which, the control point correspondences are established and a multilayer perceptron neural network is trained to refine the nonlinear warping.

We then tested the proposed mPAR method on the three-dimensional prostate models reconstructed from real surgical specimens. Such data set provides us a perfect testing case since the prostate models contain multiple internal objects and possess natural nonlinear deformations. Without knowing in anyway the control point (i.e., the contours of multiple anatomical objects) correspondences, we used mPAR to map the floating model (middle) to the generic model (left), as shown in Fig. 3. In particular, one of the research aims here was to map detected tumors into a generic model so as to establish the heterogeneity statistics of localized prostate cancer. The results shown in Fig. 3 (right) were very promising in that the tumor distributions closely resemble the measured heterogeneity and agree with the visual inspection by a senior pathologist.

To illustrate the role of the mPAR method as an effective initial step for a more refined follow-up registration, we consider a hybrid algorithm to register a sequence of mammograms for breast cancer detection. In this experiment, the cross points between vertical and horizontal elongated structures are used as potential control points. These elongated structures represent blood vessels and milk ducts. The potential control point clus-

ters are first aligned using the mPAR method to facilitate the formation of control point correspondences, and the registration is refined by a trained multilayer perceptron neural network based on the outcome of mPAR.

In this experiment, both PAR and mPAR are used to perform the initial registration which should be able to correct most of the global distortion and misalignment between the two images. The control point correspondence is then obtained by overlaying the potential control points from the new image with the potential control points of the old image and then using a nearest neighboring principle.

The raw image sequence is given in Fig. 4 and is composed of the scans of a patient acquired on (a) 3/5/96 and (b) 2/24/99. The final warped image is shown in Fig. 4(c). From visual inspection, we see that most of the scale difference between the images has been corrected, as shown by the difference image in Fig. 4(d). In this example, control point matching using a nearest neighboring method yielded 27 control point pairs out of a pool of 66 potential control points, evenly distributed across the image. This yields a match rate of 40.9%, as a relative and indirect measure of the performance of mPAR as an initial reg-

istration step. Given the difficulty of the task, our method performs relatively well in that the information on control point correspondences is not in anyway available and many of the potential control points may not indeed form pairs in fact.

IV. CONCLUSIONS

We have presented theoretic evidence which shows that principal axes registration of two point sets based upon information theory, without needing to establish explicit point correspondences, is optimum under a rigidity assumption. We have proposed a mixture of principal axes registration method, supported by a standard finite normal mixture modeling of control point clusters, for nonrigid cases. The corresponding results clearly indicate that such multiple transformational methods, in a broad sense, outperform conventional, using single transformation methods. The new methods presented in this paper would be most suitable as an initial step in other, more sophisticated image registration algorithms.

APPENDIX

Examples of Gradients [11, pp. 60–61]:

$$\frac{\partial \log |\mathbf{W}|}{\partial \mathbf{W}} = (\mathbf{W}^t)^{-1} \quad (19)$$

$$\frac{\partial \mathbf{x}^t \mathbf{W} \mathbf{x}}{\partial \mathbf{W}} = \mathbf{x} \mathbf{x}^t. \quad (20)$$

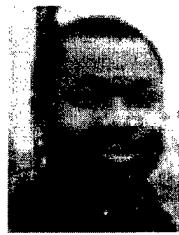
REFERENCES

- [1] M. Moshfeghi and H. Rusinek, "Three-dimensional registration of multimodality medical images using the principal axes techniques," *Philips J. Res.*, vol. 47, no. 2, pp. 81–97, 1992.
- [2] V. Govindu and C. Shekhar, "Alignment using distributions of local geometric properties," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 1031–1043, Oct. 1999.
- [3] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 698–700, Sept. 1987.
- [4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [5] Y. Wang, S.-H. Lin, H. Li, and S.-Y. Kung, "Data mapping by probabilistic modular networks and information theoretic criteria," *IEEE Trans. Signal Processing*, vol. 46, no. 12, pp. 3378–3397, Dec. 1998.
- [6] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. Berlin, Germany: Springer, 1994.
- [7] G. E. Hinton, P. Dayan, and M. Revow, "Modeling the manifolds of images of handwritten digits," *IEEE Trans. Neural Networks*, vol. 8, no. 1, pp. 65–74, Jan. 1997.
- [8] Y. Wang, L. Luo, M. T. Freedman, and S.-Y. Kung, "Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization," *IEEE Trans. Neural Networks*, vol. 11, pp. 625–636, May 2000.
- [9] Y. Wang, T. Adali, S.-Y. Kung, and Z. Szabo, "Quantification and segmentation of brain tissues from MR images: A probabilistic neural network approach," *IEEE Trans. Image Processing*, vol. 7, pp. 1165–1181, Aug. 1998.
- [10] A. D. J. Cross and E. R. Hancock, "Graph matching with a dual-step EM algorithm," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1236–1253, Nov. 1998.
- [11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.



Yue Wang received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 1995.

He is currently an Associate Professor of electrical engineering and computer science at The Catholic University of America, Washington, DC. He is also affiliated with The Johns Hopkins Medical Institutions, Baltimore, MD, as an Adjunct Associate Professor of radiology. His research interests focus on computational bioinformatics and molecular imaging.



Kelvin Woods received the B.S., M.S., and Ph.D. degrees in electrical engineering from Tuskegee Institute, Tuskegee Institute, AL, in 1991, The Pennsylvania State University, University Park, in 1993, and The Catholic University of America, Washington, DC, in 2001, respectively.

Since 1993 he has been an Electrical Engineer with the U.S. Government. His research interests include communications theory, signal processing, image processing, and parameter detection and estimation.



Maxine McClain received the B.S. degree in biomedical engineering in 1998 and the M.S. degree in electrical engineering in 2000, both from The Catholic University of America, Washington, DC. Her M.S. thesis topic was medical image segmentation and registration.

She is currently a Postgraduate Researcher at Oak Ridge National Laboratory, Oak Ridge, TN.

Ms. McClain is the recipient of a 1998 Claire Booth Luce Fellowship.

Technical Report

Yue Wang and Kelvin Woods

The Catholic University of America TR-CUA001

Chapter 1

Introduction

1.1 Background

Breast cancer is one of the leading causes of death among women today. To help combat this problem doctors use medical imaging (mammography) as a mechanism to screen patients and identify cases where further analysis is required. In breast cancer diagnosis, the mammography has proven to be the only way to detect cancer at its earliest stages, thus improving the patient survival probability[4]. A patient's survival probability is directly linked to tumor size upon detection. Tumor size has an apparent relationship to tumor grade or disease progression which can dictate treatment options. Studies have shown that women at age 40 and up are most at risk for developing breast cancer. Although this factor alone is not the sole contributor, most women over 40 have screening mammograms performed periodically (usually one or two years apart) in an effort to detect the existence or onset of a cancerous condition in the breast. This type of study is called breast cancer screening and usually is limited to asymptomatic women where craniocaudal (CC) and mediatorial oblique (MLO) mammographic views are acquired and analyzed for signs of cancer[4]. These images are reviewed manually by a radiologist following a prescribed procedure which specific viewing apparatus, lighting requirements, and amount of time per case [4]. Generally, a radiologist reviews four images of a single view (either CC or MLO) simultaneously. The images are the current left and right breast aligned over top of the left and right breast taken previously. Figure 1.1 shows the layout for the screening case. By aligning the images in this manner, change (tissue change) over time can better be identified. This tissue is a key indicator to the onset of a cancerous condition. Studies have shown a correspondence between tissue change and underlying biological change. This change is important for applications such as treatment monitoring and lesion diagnosis. Once change has been detected, further analysis of the region is performed.

1.2 Statement of Problem

Due to limited resources, radiologist often must review a massive number of cases during a period. Also, the constraints on resources have caused radiologist with less experience in mammography analysis to review cases. The review of this massive volume (around 8 images per case) of data and inexperience could cause missed tumors, delayed detection, and false positives which ultimately cause a reduced life expectation upon detection, unnecessary patient call backs, and unneeded needle biopsies.

To reduce some of the load on the radiologist and to improve diagnosis accuracy, development of automatic computer aided diagnosis (CAD) system for change detection have been explored [5], [6], [69]. These systems aim to automate portions of the analysis process. In order to accomplish this task, one must roughly model the analysis task performed by the radiologists in the course of an examination. Since this research focuses on change detection, the task modeling discussed here focuses on that task. The radiologists's analysis process consists of the following steps: (1) Acquire mammograms of previous and current visit; (2) Mount the image in specific order (see Figure 1.1); (3) Mentally examine images for similar landmarks and mentally adjust view; (4) Identifying corresponding regions and compare for change. From the examination of these four tasks, it is apparent that steps three and four would stand to benefit the most from automation as steps one and two are relatively simple.

Several key issues make automation of steps three and four extremely difficult, with step three being the most difficult. The issue is the fact that mammograms are complex images that do not contain any clearly defined landmarks. Secondly, differences in breast positioning and compression during acquisition could cause images of one scene to visually appear different. Finally, breast sizes and consistency can vary with time (e.g. weight loss, surgery, and age). The research of the clinical problem of change detection in a mammogram sequence of a single patient uncovers serval difficulties and complex technical problems. The first problem is how do you align a generally non-rigid object without apparent control points or landmarks? This problem is classified as a image registration

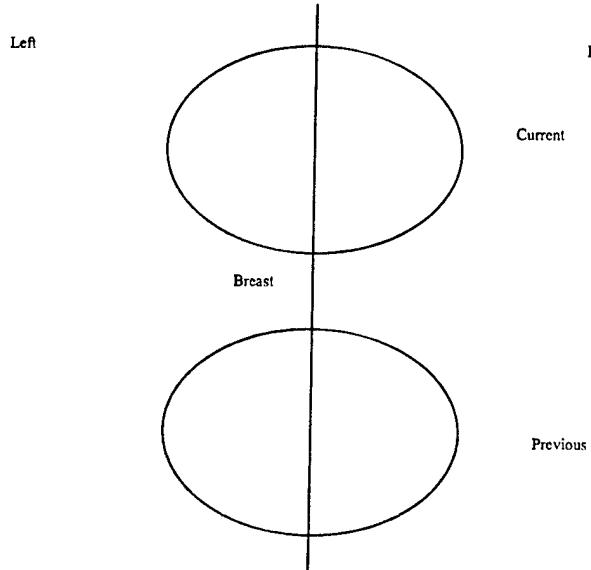


Figure 1.1: Layout of Screen mammogram analysis

problem. Image registration has been the topic of much research over the years [73]. The other problems are directly related to change detection. In mammograms, drastically different images can be attained from the same patient imaged at temporal displayed times. The key questions here are how do we discriminate natural change from cancerous change and how do we determine the type of change that has occurred? Often in medical imaging, the type of change that has occurred can direct the type of treatment required. Prime examples are treatment monitoring and tumor detections. This process we define as change quantification. This definition was motivated by the work of [18] where quantification is used to define the process of describing the image with some model parameters. The specific aim of this research is to study image registration and change detection to address the clinical and technical problems discussed above. The result will be a semi-automatic change detection algorithm.

1.3 Technical Review

Two main approaches were developed to deal with the problem of automatic change detection in mammograms. They are approaches based on processing a single view of a single breast [5], [6] and approaches based on single view of multiple breasts (left and right) [8]. [69] presented work that developed an approach to consider both single and multiple view processing. Use of multiple breast views leads to additional problems because women typically have significantly different structures between left and right breasts [1]. This causes natural asymmetry to be flagged as change or lead to landmark confusion [5] while single breast approaches do not have the problem of dealing with asymmetry. So, most of the research attention has been focused on single breast approaches. Generally, single breast approaches contain three main steps: (1) preprocessing of the images searching for control points or regions for use in registration, (2) registration, to align the images into a common framework, and (3) detection and analysis of local change. The preprocessing is generally handled by classical image processing techniques such as segmentation, morphological filtering, edge detection, and feature extraction. The registration process is performed by both rigid and non-rigid forms, but generally the breast is considered a deformable object thus non-rigid forms of registration should be used [73]. Finally, the local change analysis is performed with various techniques ranging in complexity from difference image analysis [15] to principle component analysis [81].

Three main research groups have attempted to address the problems of mammogram registration and change detection. Group [5] approached these problems with a two layered approach. In their approach, they perform a sequence of two polynomial based (thin-plate spline TPS) registration using different sets of control points. The first set of control points were extracted from the smoothed dense tissue boundary (i.e. brightest region on mammogram). The second set was extracted from the interior region of the dense tissue. Correspondence between control points for the first transform was performed by matching points on the reference image contour with similar points on the float image contour with the same maximum curvature. For the second transform, points with matching LAWS's texture features [87] were matched as control points. This approach has problems when the dense tissue does not

occupy a large percentage of the breast which typically occurs in radio-lucent breast [1, Breast book]. In cases like this, error occurs in transforms when the point to be transformed is far away from the control points thus reducing the effectiveness of the control points.

Another approach to mammogram registration and change detection was developed by [6]. They consider these problems by asserting that accurate registration of mammograms is intractable except with elastic transforms, and the only solution is regional registration [7]. In regional registration, localized areas of the two mammograms are aligned based on their distance from control points. In their approach, monotony operators are used to extract vertical and horizontal elongated structures (milk ducts, and blood vessels) in the image which they assume to be generally stable between images in the sequence. A three-pass Gaussian filter is used on the original mammogram to mask less prominent structures. This reduces the complexity and limits the monotony operators to detecting the dominate structures. The cross points of these horizontal and vertical structures make up the pool of potential control points. Correspondence between the current image control points and reference image control points is accomplished by comparing the respective control point signatures. The signatures are created by counting the number of non-zero pixels that lie in a rectangle that is rotated around the control point. In this configuration, the direction of the longest structure would yield the highest value in the signature. The similarity of the signatures is used as the matching criteria. These values are then passed into a thresholded accumulator matrix for final point selection. To localize the area where signatures are compared, the nipple location in both images is used to determine a neighborhood region that surrounds the potential control point. This reduces processing and decreases the probability of false alarm. Using these control points, regions (of any shape) are determined on the current image by calculating the distance from a subset of the detected control points. Finally, the regions are compared for change. This method overcomes the erroneous interpolation problem experienced by [5], but the algorithm uses ad hoc point matching criteria, localize window size selection, and threshold determination. In addition, [7] assumes a small mis-registration that restricts the generality of this approach. Both [5] and [6] mainly address registration so, simple change detection methodologies based on difference image analysis and wavelets respectively are used for their change analysis. In [9]'s approach, the registration is performed by a radial basis function (RBF) interpolation process. This approach as other in polynomial based registration methods depends heavily on the existence of control points in the image pair. This approach only uses control points on the skin line of the breast which has been extracted through threshold based image segmentation. Control point correspondence is obtained by finding contour points that are equidistant (measured in the number of contour points from the corresponding nipple) from the nipple. The control points are then used to solve for RBF parameters which yield the desired transform. Since the control point are selected only from the skin line, internal structures are not considered in registration. Thus, this method is unable to track non-rigid changes that occur inside the breast. In addition, use of threshold based segmentation could lead to a noisy contour.

Although these methods have had success on limited databases, their limitations could cause erroneous results when examining mammograms in a more general sense. For instance, consider a mammogram sequence where both images contain a small dense tissue area (relative to total breast tissue size). Using [63], the control points would be clustered around the dense tissue area leaving the rest of the image not modeled. So, any transform derived from these points could not accurately capture any deformation in the not modeled portion of the image thus causing mis-registration. In addition, consider that the same sequence has a large initial misalignment. This causes the window sizes, thresholds, and signature matching criteria of [6] to be manually modified to correctly process. The approach [69] is insensitive to the above conditions, but would not accurately model the internal structures because no control points exist in that region. This short fall could possibly cause the detection of false or missed change. The limitations of [5] [6] [69] are listed in Table 1.1.

Another problem not considered by the above three approaches is a sequences containing more than two images (i.e. I_i , I_{i-1} , I_{i-2}, \dots). Sometimes in medical analysis, the radiologist will examine further back than previous images as some change can only be seen over a longer periods of time. In satellite imaging, site monitoring is a similar task. In this task, sites are monitored through several images (generally two or more). To accomplish this task [79] uses the site model. The site model is a multimedia representation of an image scene to include object shapes location, segmented version of scene, previous location of change, extracted features, and a prior domain expert information. Through the site model operations of construction, registration, and update the site model tracks the scene over time. This same approach could be used to analyze an anatomical region such as the breast, brain, or prostate in temporal studies.

1.4 Approach

Thus, considering the limitations listed in Table 1.1 and site model theories, a new algorithm is proposed to perform non-rigid registration applied to a mammogram sequence. In this algorithm the registration is perform in two steps. The first step is called initial registration and it aims to correct large global misalignment by treating the breast as a sum of rigid objects and performing a multi-object principle axis registration(PAR). The objects include large

Limitations	Effect of Limitations
Wirth Method	
Only use control points on the skin-line.	Unable to consider deformation of internal structures
Number of contour points between control points as measure of control point matching.	Assumes that the number contour points between two control points is constant across the float and reference image.
Difference image analysis (detection only).	No quantification
Sallam Method	
Used the boundary and interior of dense tissue to determine control points.	Control points do not model complete image deformation in case when dense tissue is a small percentage of image
Used threshold methods to segment image.	Yields different contours if intensity ranges differ for reference and float image.
Difference image analysis (detection only)	No quantification
Brzakovic Method	
Assume small initial mis-registration.	Limits use to cases of small registration.
Image dependent processing parameters such as signature search window, size of monotony operators, and thresholds.	Requires new parameters for each image.
Histogram analysis using raw images (detection only).	No quantification
Adhoc signature matching method.	Assumes the longest arm of signature will remain the same in float and reference images.

Table 1.1: Limitation of existing Mammogram registration algorithms

clustering of similar tissue types and the breast skin line. An individual PAR transform is calculated for each object. Each pixel x_i is then passed through each of the T_k transforms resulting in multiple point matching \hat{x}_{ik} in the new image. The final point location \hat{x}_i is formed by weighting each point \hat{x}_{ik} by the probability z_{ik} that the point x_i was transformed by T_k (or probability that x_i belongs to class k). z_{ik} is derived by considering each of the objects as a cluster of control points described by a normal distribution. Thus similar [19], we assume that each (x, y) locations to be made up of a sum of these normal distribution which can be modeled as a finite mixture.

This formulation allows for a weighting of the transform T_k to determine the final transform T . Thus, creating a global interpolative transform that weights local characteristics based on their probability of membership. The next step in the registration process is called final registration. In this step, non-rigid displacements between images are accounted for using a polynomial based (thin-plate spline) registration. Polynomial based algorithms depend heavily on the existence of control points between the images. To obtain the control points, we follow a modified version of the approach discussed in [7] which it extracts the elongated structure from the mammogram and uses the cross points of vertical and horizontal structures as the control points. The approach is modified by using the Pearson correlation coefficient [14] to match the potential control point signatures instead of the direction of the longest arm of the signature.

Similar to registration, change detection is performed by a two step process. The process consists of a detection phase and quantification phase. The detection phase consists of measuring the relative entropy between the joint histogram of the float and reference images with the joint histogram of reference with itself. The quantification phase uses basic geometry to determine an object's area and center of gravity which are then compared to determine if the object has change. To add the ability to study longer sequences, the site model was used to support the registration and change detection process. The site model supports the registration process by defining a reference frame which all subsequent images will be registered. The site model also fuses user input knowledge with automatically extracted data into a single model to be used in the registration process. As for change detection the site model stores the detected changes along with site memory and any other parameter updates.

The automatic change detection algorithm can be summarized into three main steps as outlined below.

Initial Registration

Preprocess mammogram for skin line and internal objects.

Use multi-object PAR on breast tissue using the skin line and internal object to form a finite cluster transform.

Final Registration

Preprocess the PAR transformed image searching for control points and transform coefficients.

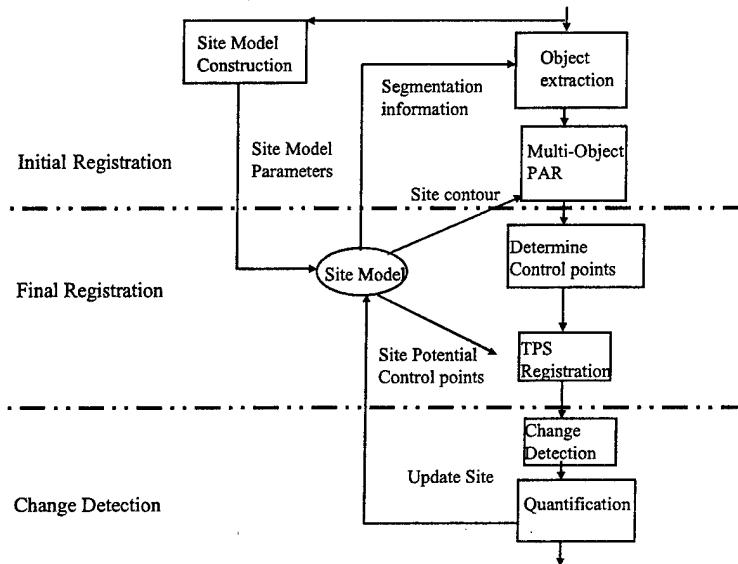


Figure 1.2: Change detection processing system flow

Use TPS formulation to determine the required transform.

Change Analysis

Use relative entropy for change detection criteria between the image blocks.

Quantify change by determining difference in object area and center of gravity.

Update change map located in the site model.

A complete flow diagram of the process is shown in Figure 1.2.

1.5 Research Scope

During the development of this algorithm, several assumptions were made in order to bound the scope of this research. First, the mammograms are assumed to be CC and MLO views only (i.e. screening mammograms) of the same patient acquired overtime. Second, the radiologist initializes the site model parameters by identifying areas of interest (local change windows) and other prominent landmark points (calcifications, large blood vessels) in the first image of the sequence. Third, the type of change was limited to growth of a mass, or shrinkage of a mass. Microcalcification changes can be detected, but will not be considered because drastic gray level difference between microcalcifications and non microcalcifications. Although, if present in both images of the sequence they may be used as control points. Fourth, the amount of initial mis-registration is bounded so the skin lines of each breast are not more than $\pm 25^\circ$ rotated from each other.

1.6 Contributions

The pursuit of this research has led to several contributions in image processing and medical imaging. Contribution one is the development of a new hybrid registration algorithm aimed at the registration of non-rigid objects with minimal a priori knowledge. Usually, non-rigid objects are registered with elastic or deformable methods which require knowledge of a sufficient number of control point pairs. While some rigid methods relax this requirement and usually only require object correspondence, for example, surface matching and principle axis methods. Use of rigid methods alone, in non-rigid problems, would allow for limited correspondence knowledge, but could not accurately model expected non-rigid deformations. The purposed algorithm combines rigid and non-rigid techniques to accomplish the registration tasks. The algorithm consists of two steps an initial step (rigid transform) which preforms multi-object PAR registration where object correspondence is assumed known, and a final step (non-rigid transform) that uses thin-plate spline (TPS) based mapping where control point correspondence is determined via a detection and correspondence algorithm. The combination of these two steps is new and provides many advantages over existing methods. The first advantage is no requirement for point correspondence in the initial step. Only object

correspondence is required which is usually much easier computationally to determine. True, point correspondence is required at some point in the processing, but performing the determination after the image has been preliminarily aligned should allow for a more focused or narrow control point search windows because potential control points should now be closer spatially. The second advantage is the ability to model non-rigid transforms by considering each rigid transform as a piece wise component of a total non-rigid transform similar to modeling a non-linear function by linear pieces[77]. This approach is a departure from traditionally registration approaches which usually follow either rigid or non-rigid transforms[73].

Contribution two is the development of a new change metric based on the joint relative entropy between two images. Unlike other change detection metrics [10], the joint relative entropy is useful in detecting translation only changes. In addition, the result of the metric tell us how similar the blocks are to each other. Difference image analysis is also useful for translation change, but it is highly sensitive to noise and does not yield a measure of how close the blocks of data are to each other.

Contribution three is the application of the site model concept to medical imaging. The site model was developed to monitor a site from a sequence of aerial images [13]. In medical imaging, the site model idea was modified to accomplish application such as lesion monitoring, and disease detection. In addition, through update procedures the site model allows for the examination of the entire sequence together, to show region progression or to further highlight small changes. The main modification to the site model idea was the creation of another variable to store changes. In traditional site model formulations, new objects are added back into the image, but in the medical environment the site image is untouched. The changes are stored in the change map. The site image is untouched because it forms the base frame for comparison so any modification could alter results.

Contribution four is the development of a methodology to combine multiple transforms together to determine a composite image transform. In this research, we apply the combination method to multiple PAR transforms, but the method is generic and can be applied to any type of transform along as each cluster control point meets the particular requirement of the registration method in question. For example, to use an elastic registration method it is assumed we know the point correspondence of control points. In this algorithm, the image is assumed to contain several clustered control points, which follow a normal distribution, for which cluster correspondence is known (i.e. objects). The resulting transform now enables rigid transform methods to handle non-rigid transform assuming the clusters are sufficiently distributed throughout image.

Contribution Five is the development of a new statistical segmentation algorithm for sequences of images. This algorithm is used in the site model update to reprocess the segmented image given the images of the sequence. The major assumption is that the adjacent images contain the same view. This algorithm is based on a 2D statistical segmentation algorithm where pixel relationship is assumed across adjacent pixels in the (x, y) direction. The algorithm extension takes advantage of the relationship between adjacent images. So, pixel neighborhood is considered in three directions (x, y, z) . This additional information leads to a more robust segmentation as seen in [54].

1.7 Report Organization

This report is organized into seven chapters. The first chapter contains an introduction, background, problem statement, and contributions. The second chapter gives a brief tutorial on mammogram formation and screening procedures. Chapter three discusses the algorithms involved in the site model construction and update. Followed by chapter four that contains the techniques for image-to-site model registration. Chapter five discusses change detection while chapter six presents and discuss global results. Finally, chapter seven presents future research direction.

Chapter 2

Mammography formation and Screening

2.1 Introduction

Breast cancer is one of the leading causes of cancer related deaths among women. Each year more than 100,000 cases are diagnosed and more than 40,000 women die[1]. For many years researchers have studied breast cancer in search of an understand of breast cancer development. A high prediction rate of who will develop breast cancer is still an impossible task; although several factors have been identified as leading to the increase risk of breast cancer development. These factors include: gender, age, family history, age of first-term pregnancy, and previous history of breast cancer. Because of the gender factor, all women are at risk of developing breast cancer. In fact, women are 100 times more likely of developing breast cancer than men [4]. Breast cancer is a progressive disease, evolving through stages of growth. The size of the tumor size when detect has an apparent relationship to tumor grade and should be considered an important prognostic factor. Mammography, a form of X-ray imaging, has been shown to be the only method currently available for the reliable detection of early, non-palable, and potentially curable breast cancer [3]. So, women starting around the age of 40 are imaged every two years or so. These mammograms are put through rigorous examination for possible cancerous regions utilizing a process called screening mammogram. The rest of this chapter is organized as follows: tutorial on mammogram formations, and explanation of screening mammogram process.

2.2 Mammogram Formation

Mammography is an X-ray image of the breast used to detect, diagnose, or monitor cancerous conditions. It is usually performed by a trained technician with the ultimate goal of imaging as much breast tissue as possible. The patient is usually standing with her breast compressed against a support plate [2]. Compression of the breast is performed to equalize the thickness across the breast which produces a uniform image. A mammogram system is generally composed of four main components: X-Ray generator, compression device, scatter grid, and acquisition hardware. The general mammogram process is defined by these four steps. (1) arrange the breast in the compression apparatus. (2) Transmit a given X-Ray spectrum through the tissue. (3) Collect the X-rays and calculate the signal strength. (4) Form image using the results from step (3). Figure 2.1 shows the arrangement of the components in relation to the breast to be imaged. The usability of the images is directly dependent on the image quality. Image quality is effected by several interrelated factors such as: contrast, which is useful in soft tissue examination; unsharpness, which is useful for small calcification; amount of X-Rays absorbed by breast tissue, where higher level increase contrast but put the patient at risk for radiation-induced carcinogenesis [4]; and high dynamic range which handles variation of the transmission over the entire mammogram. Thus, the goal is to determine compromises that best match the given factors. Next, each of the components in the Figure 2.1 will be discussed in more detail.

X-rays are produced by energy conversion when high speed electrons from the cathode hit the anode target as shown in Figure 2.2. The electrons are discharged from the cathode as a result of heating. This discharge is called therminic emission. X-rays (photons) are created when the electrons hit the atoms present in the anode. The area of the anode that is bombarded by the electrons is called the focal spot. The focal spot is directly related to image resolution. The smaller the focal spot the better the resolution. Since the X-ray emission from the anode is isotropic, shielding is needed to reduce undesired exposures to the patient and film. The shielding is performed by an elongated tube with a single opening. The tube opening is capped with a collimator to further reduce unwanted radiation emission.

The radiation is composed of three general energy levels low, medium, and high. The low and high energy photons are filtered out because low level photons usually are attenuated some much by the tissue that they do not

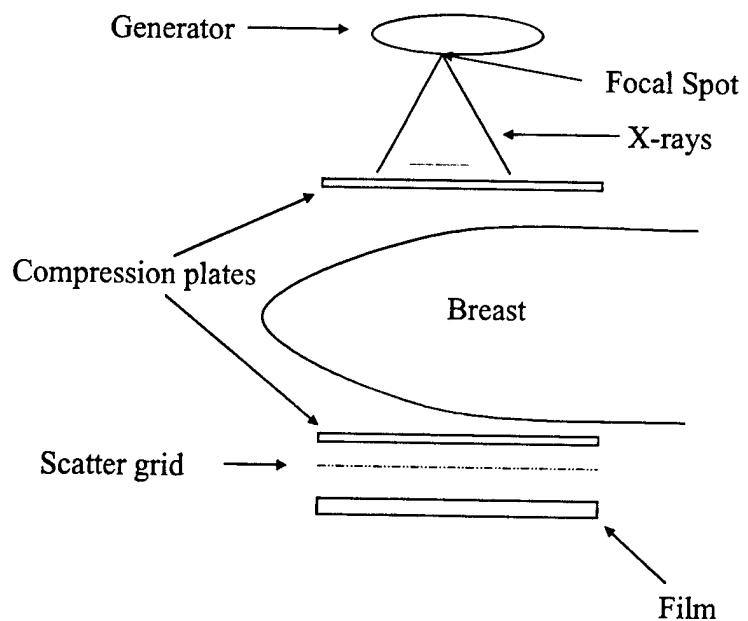


Figure 2.1: Mammogram System

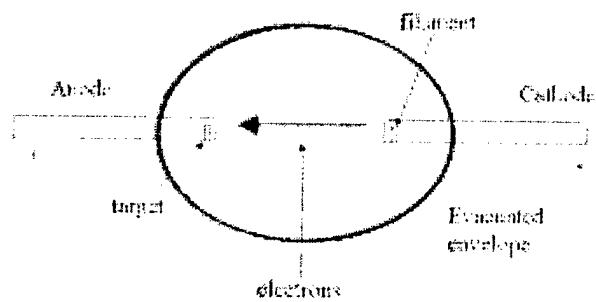


Figure 2.2: Major components of X-ray

reach the film and the high level photons are unchanged by the tissue causing a low contrast image. This filter is used to shape the spectrum to achieve the best image quality. Most frequently, a molybdenum filter is used, but this is variable based on breast composition and thickness. Breast tissue composition goes through several phases of development during a women's life. In each of these stages the breast can be composed of different tissue types. For example, in infancy the breast is mostly composed of adipose tissue while in puberty the fibroglandular tissue develops, and in maturity the fibroglandular tissue is replaced by fat tissue. Each of these tissue types attenuate the X-rays differently which yields different absorption rates.

The next component of a mammogram system is breast compression. Compression is performed using flat compression plates. A main advantage to compression is the breast tissue is forced to a uniform thickness. This avoids the problem of overexposing the thinner regions (near nipple) and underexposing the thicker regions (near chest wall). A second advantage is that the compression holds the breast in place during imaging. This reduces image unsharpness caused from tissue motion. Other advantages of compression are reduced absorption rates because the breast tissue is now thinner, shorter exposure time because the x-ray have a shorter distance to travel, and confusing and overlapping structures are separated.

Following the breast compression is the scatter grid. The scatter grid is designed to drastically attenuate the photons that are hitting the plate obliquely. These photons are more than likely the result of scattering from within the breast tissue. Scatter grids are composed of thin strips of metal laid with a particular spacing. Grids come in variety of different configurations. They are measured using a term called grid ratio. This is defined as the ratio of the length to strip spacing. When the scattered photons are removed there is an increase in the image contrast. In [2] contrast was improved by 17%, 37%, and 54% with the use of filters with ratio values of 2,4, and 8.

The final component of Figure 2.1 is acquisition hardware. Acquisition hardware includes the process that receives the photons from the scatter grid and then translates it onto the film. This process contains two major steps. The first step converts the photon into visible spectrum by exposing a luminescent intensifying screen to the photons. This reaction produces light which is then used to expose film and form the radiographic image. Next, this image is transformed into a visible image by standard developing techniques.

2.3 Mammogram Screening

Screening mammograms is the term given to the periodic mammograms used in early detection of possible cancer conditions. The question the radiologist wants to answer using mammograms is, "Is this mammogram completely normal or is additional analysis required?" The major goal of mammography is to image the breast in order to detect cancerous conditions at its earliest stages. With this goal in mind technicians generally try to arrange the breast to image as much of the tissue as possible. Since the breast is a three dimensional organ, it is important to obtain multiple views so confusing or overlapping structures can be resolved. Generally, in screening studies the mediolateral oblique (MLO) and craniocaudal (CC) projects are obtained [1]. Together these two projections visualize the majority of the breast tissue, although, if sufficient compression is not achieved then the deep tissue close to the chest wall will not be imaged. Figure 2.3 and Figure 2.4 shows examples of CC and MLO compression views with a corresponding mammogram.

The mediolateral oblique projection is the most useful projection because this view projects most of the breast tissue onto the image including breast tissue close to the chest wall. In this projection, the compression plane is oblique not the patient. The compression plane extends through the nipple from the upper outer quadrant of the breast to the lower inner quadrant of the breast as shown in Figure 2.4. On the other hand, in the craniocaudal projection the compression plane is perpendicular to the chest wall. This view shows the thinner portion of the breast, but can often miss the thicker portion because of positioning. Usually, after the MLO and CC views have been examined, additional views may be required depending on the review results. The other supplement views include: lateral, medial, lateromedial, and straight mediolateral. Use of these views depends heavily on the particular cancerous sign.

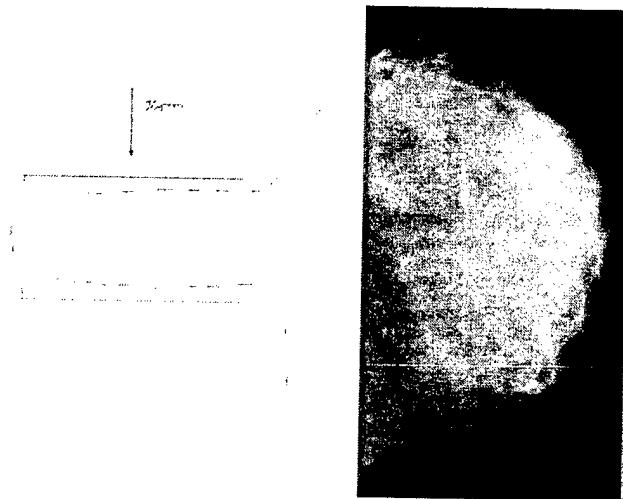


Figure 2.3: Compression plain and sample CC mammogram view

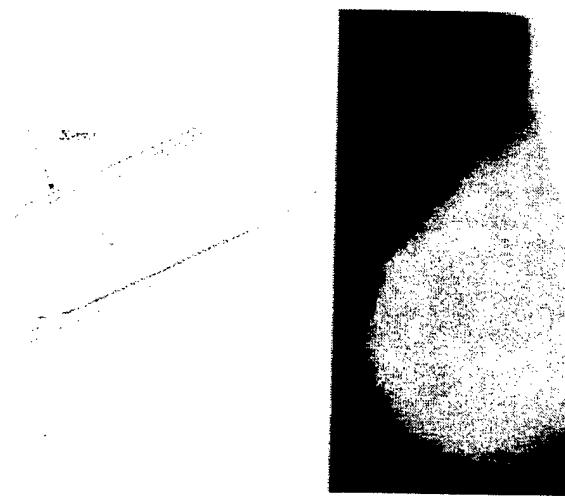


Figure 2.4: Compression plain and sample mammogram for MLO view

Chapter 3

Patient Site model Construction and Update

3.1 Introduction

The site model is a dynamic mathematical and geometrical description of a scene under analysis. At a minimum, the site model contains the following parameters: objects, boundaries, object attributes, user input, and associated raw and processed data. The site model can vary in complexity ranging from detailed object description (building numbers) to simple boundary information. Pioneering work on the site model was performed by [13] in the analysis of aerial images for site monitoring and change detection for intelligence gathering purposes. In that research, the goals of site monitoring and change detection were accomplished through the support of three main model tasks. These tasks are called site model operations. In [79] the operations are defined as site model construction, image-to-site model registration, and site model update. Other research on the site model idea was performed by [74]. In [75], the site model operations are defined as site model acquisition, model-to-image registration and model extension. The pursuit of both of these research projects resulted in algorithms for automatic building detection [13], automatic and semi-automatic registration [79], [75], and fusion methodologies for combining user input with automatic processing results. Next, each of the site model operations will be further defined and discussed.

The first model operations is site model construction. Site model construction consists of deriving the site model parameters from the initial input images and user input. In [79], the construction process is as follows: (1) review two or more input images (overlapping views); (2) create a world coordinate system; (3) derive camera models for each image; (4) input camera focal length and principal point; (5) determine control points; (6) refine camera models for each image; (7) add objects and other annotations. [75] on the other hand, considers a lower level construction phase which includes (1) line segment extraction, (2) building detection, (3) multi-image epipolar matching, (4) multi-image triangulation, and (5) projective intensity mapping. These site model parameters which include detected line segments, buildings locations, camera models, and other control points are extracted using advanced and classical image processing techniques.

The next site model operation is image-to-site model registration. Image-to-site model registration is the process of putting a new incoming image (float image) into the same coordinate system as the site model (reference image). The registration process may be automatic or semi-automatic (user interaction). A general approach is to match, in some manner (via. criteria), selected site model parameters with newly extracted parameters in order to derive a transform that describes the recovering transformational geometry (transform) required for alignment. [79], [74], [78] describe several registration methods that they use with their site model. The result of this operation is an aligned image ready for change analysis.

The site model's ability to describe a scene over time is derived through the site model update procedure. Site model update allows for the addition of parameters (objects) of the site based on processing results of previous and current imaging conditions. With these operations, site change, such as a vehicle leaves a parking lot or lesion increase in size, can be detected and monitored efficiently. To maintain continuity, [79]'s notation for site model operation will be used throughout the rest of this report.

The site model idea can be extended to medical imaging analysis. In medical imaging, the radiologist often wants to perform similar types of applications to site monitor and change detection. For example, lesion detection and treatment monitoring. In these applications, a radiologist examines a temporal sequence (same view) of the same patient for change that could indicate cancer. When change is found further analysis is performed. For example, in mammogram screening, temporal sequences of the same patient are used to detect possible regions of interest.

Currently in medical imaging another type of model is used in various processing algorithms [52] called anatomical atlas (models). Although anatomical models are currently not used in change detection application, it is important

	Parameter	Size
1	Skin line	$2 \times N$
2	Raw image	$M \times M$
3	Segmented image	$M \times M$
4	Mask	$M \times M$
5	Center Gavity	1×2
6	Eigenvalue	2×2
7	Eigenvector	2×2
8	Nipple location	1×2
9	Elongated Structures	$M \times M$
10	Potentital Control points	$n \times 2$
11	Image histogram	$1 \times M_{gl}$
12	change map	$M \times M$
13	Internal objects	$k \times 2 \times g$
14	Control point Signatures	$\frac{N}{366} \times n$
15	Quantification parameters	$K \times 3$

Table 3.1: Site model parameters

to note the differences between the anatomical model and the site model. The main difference between the site and an atlas is the site model is specific to a particular scene (patient) where an anatomical model is more a textbook rendering of the scene that does not consider user input or individual variability. An example is an anatomical atlas of a MRI brain [57]. In this example, the synthetic brain MR image has the correct tissue percentages. This difference leads to a more refined name for the site model called the patient specific site model.

In this research, the site model is used to support registration and change detection to achieve the application goals of lesion detection and treatment monitoring in mammograms. The site model supports registration by providing a common frame (coordinate system) from which all other images in the sequence are registered. It also provides an efficient mechanism for combining manual site information (user label objects) and automatic information (detected boundaries and control points) in a useful manner to help facilitate the desired task. The rest of this chapter considers the specific contents of the model, the signal and image processing techniques used to construct the model parameters, and the site model update procedures.

3.2 Model Parameters

In this section, the site model components will be listed and their relevance discussed. Since the site model will be used to support sequence registration and change detection, it contains parameters used in the accomplishment of these tasks. Parameter order in the site model is arbitrary as the site model is interactive and parameters are used in a non-linear fashion. There are some parameters that depend on others, and naturally the dependent parameters would need to be calculated after the required information was available. The site model parameters included in this implementation are shown in Table 3.1. Next, the purpose of each parameter is discussed.

The first parameter is a $N \times 2$ vector containing the x, y coordinates of the breast skin line. The breast skin line parameter is used in initial registration as one of the multiple control objects and as the desired curve to be fit in nipple location estimation. The second parameter in the model is an $N \times 2$ vector containing the (x, y) coordinates of the largest objects, usually dense tissue, located within the breast tissue. These objects are used in conjunction with the skin line to perform multi-object registration. The third parameter is the $N \times 2$ contain the x, y locations of potential control points. These points are the cross points of horizontal and vertical structures (blood vessels and milk ducts) within the breast. The points are used to form the spatial-coordinate transform in the final registration phase. The fourth parameter is an image containing both horizontal and vertical structures. This image is used to generate point signatures for the determination of point correspondence between potential control points in reference (site) and float (incoming) image. The fifth parameter is the estimated nipple location and is stored in a 1×2 vector. The nipple location is used to localize point correspondence to a neighborhood window in the correspondence phase of final registration. The sixth parameter in the site model is the raw image histogram stored in a $1 \times M_{gl}$ vector (M_{gl} is the maximum intensity value in the image). The histogram will be used as the desired histogram in performing histogram specification between the incoming image and site. Histogram specification normalizes intensity ranges to that of the site model so object extraction is not biased by intensity differences. The eighth parameter is the image quantification model parameter estimates. These estimates are used

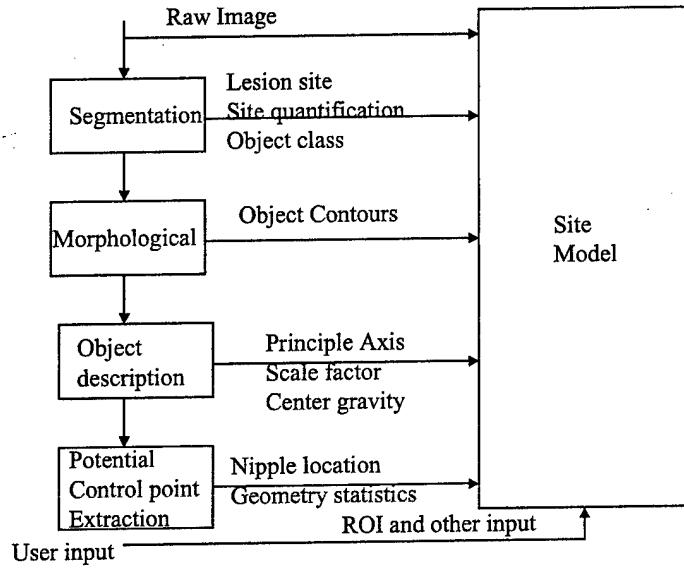


Figure 3.1: Site model construction flow

to initialize the segmentation of the incoming images so a uniform segmentation is achieved between the images. A copy of the raw image, raw segmented image and tissue mask are included and used in follow-on processing. Then, finally space is assigned for user specific input; such as the, number of classes in the scene, prominent landmark locations, change region of interest, and location of previous change. The number of classes in the scene is used to initialize the segmentation process. Prominent landmarks provide addition control points in final registration. The previous change location is used to exclude the change regions from further processing or focus in on specific regions for analysis.

The site model construction process is summarized in Figure 3.1. See Figure 3.2 for an example of a site model of a CC view mammogram. Next, the theory and algorithmic formation of each of the parameters will be discussed.

3.3 Model Construction

3.3.1 Segmentation

The segmentation algorithm used in this research is a statistical based algorithm that classifies each pixel as belonging to one of the K classes. The main premise of this algorithm is that the image's distribution can be represented by the gray level histogram of the image. The histogram of an image is defined as the number of times a pixel intensity falls within a pre-specified range as shown below.

$$p(u) = \frac{1}{N^2} \sum_{i=1}^{N^2} I(u, x_i) \quad (3.1)$$

$$I(u, x_i) = \begin{cases} 1, & u = x_i \\ 0, & u \neq x_i \end{cases} \quad (3.2)$$

where x is the intensity level of the pixel and I is an indicator function. Then it is assumed that the histogram can be mathematical modeled (or composed of) by a sum of K Gaussian distributions or mixture model where each individual Gaussian distribution identifies a class (tissue type). Finally, each pixel is assigned a class based on its membership probability. The algorithm is composed of two main components: quantification and segmentation. The quantification phase consists of estimating the parameters of the mixture model while the segmentation phase uses these estimates to determine pixel labels in a maximum likelihood sense.

Several studies of natural image statistics have yielded some stochastic image mixture models that best model the histogram of the X-ray mammographic images[19]. For this research we selected the standard finite normal mixture (SFNM) model as the histogram model. SFNM can be derived using the following relationships. First the

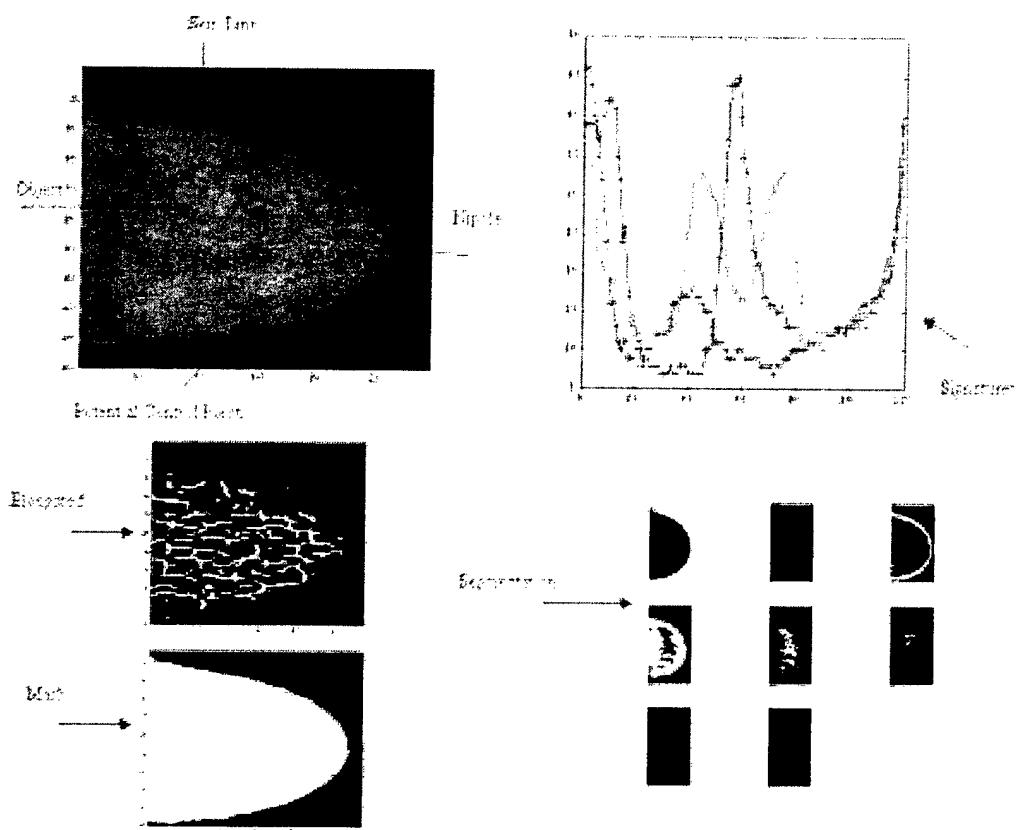


Figure 3.2: Example Site model

image is a $N \times N$ image where each pixel is assumed to be a random variable. The marginal distribution of the random variable (pixels) is shown below.

$$p(x) = \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right) \quad (3.3)$$

where x is the pixel (random variable), μ_k is the k^{th} class mean, σ_k^2 is the k^{th} class variance, and π_k is the distribution parameter. The SFNM is derived by randomly reordering the pixels with no regard to spatial information. This allows the pixels memberships to be treated as i.i.d. random variable. The joint distribution of the image is written as the product of each pixel's distribution as shown below.

$$P(X) = \prod_{i=1}^{N^2} \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \quad (3.4)$$

The above equation represents the SFNM model which can be rewritten in the form of a likelihood function conditioned on θ , the free parameters vector.

$$P(X/\theta) = \prod_{i=1}^{N^2} \sum_{k=1}^K \pi_k g(x_i) \quad (3.5)$$

$$g(x_i) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \quad (3.6)$$

$$\theta_k = \pi_k, \mu_k, \sigma_k^2 \quad (3.7)$$

In order to use this equation, the feature vector θ_k and K must be estimated. Since the components of θ are not treated like random variables, the estimation problem is formulated as a maximum likelihood (ML) estimation problem [76]. The main goal of ML estimation is to determine values for θ and K that cause X to occur. Since the logarithm is monotonically increasing, maximizing the log-likelihood is equivalent to maximizing the likelihood function [76]. The ML estimate θ' , is that value of θ that maximizes the log-likelihood function. This estimate can be determined by differentiating the log-likelihood function $\log P(X/\theta)$ and setting it equal to zero (i.e. find the extreme point of the log likelihood function).

$$\left[\frac{\delta \log P(X/\theta)}{\delta \theta} \right]_{\theta=\theta_m} = 0 \quad (3.8)$$

Sometimes maximizing $\log P(X/\theta)$ is too complex to solve in a closed form solution. In cases like this, an iterative algorithm called the expectation-maximization algorithm (EM) can be used [25] to obtain the required ML estimates. The EM algorithm is designed to attack what is termed 'incomplete data problems' [25]. Incomplete data problems are defined as problems where part of the data for some reason is unobservable. Take, for instance, the true pixel labels L of an image as unobservable data and the pixels intensity Y as observable data. The relationship between observable and unobservable data is shown below

$$X = (Y, L) \quad (3.9)$$

$$X = T(L) \quad (3.10)$$

where X is the complete data and T is a nonreversible many-to-one transformation of L . If L could be observed directly then the complete information about the image would be known and no processing would be required. The EM algorithm is divided into a E step, where the likelihood unobservable data L is calculated through the observable data Y and the current parameter estimates, and a M step, where the unobservable likelihood function is maximized to yield new parameter estimates. In the SFNM formulation, the E step, for a assumed number of class K , this is formulated as a membership functions shown below

$$z_{jk}^{(m)} = \frac{\pi_k^{(m)} g(x_j/\mu_k^{(m)}, (\sigma_k^2)^{(m)})}{f(x_j/\theta^{(m)})} \quad (3.11)$$

where m is the current iteration number ranging from 0..... Then in the M step the updated parameters (μ, σ^2, π) are calculated by maximization of the likelihood with current estimates. The update equation are shown next.

$$\pi_k^{(m+1)} = \frac{1}{N} \sum_{j=1}^N z_{jk}^{(m)} \quad (3.12)$$

$$\begin{aligned}\mu_k^{(m+1)} &= \frac{1}{N\pi_k^{m+1}} z_{jk}^{(m)} x_j \\ \sigma_k^{2(m+1)} &= \frac{1}{N\pi_k^{m+1}} \sum_{j=1}^N z_{jk}^{(m)} (x_j - \mu_k^{(m+1)})^2\end{aligned}$$

The EM iterates back and forth until a convergence criteria is reached (under regularity conditions) [25]. The convergence criteria is reached when the difference between $\pi_k^{(m)}$ and $\pi_k^{(m-1)}$ is smaller than some pre-determined value ϵ .

$$|\pi_k^{(m+1)} - \pi_k^{(m)}| < \epsilon \quad (3.13)$$

A key factor in the use of the EM algorithm is obtaining a reasonable initialization of parameter estimates[25]. If initialization is not appropriate, then the algorithm could estimate into a local minima [25]. To combat this problem, the Adaptive Lloyd Max Histogram Quantization algorithm (ALMHQ) is used to determine the initial parameters estimates for the EM algorithm. [20]. The ALMHQ algorithm takes the image intensity histogram p and number of regions K as input then iteratively determines each of the K threshold values by trying to minimize the global distortion D with respect to the thresholds t and mean gray levels μ .

$$\frac{\delta D}{\delta t_k} = \frac{\delta D}{\delta \mu_k} = 0 \quad (3.14)$$

$$D = \sum_{k=1}^K \int_{t_k}^{t_{k+1}} (\mu - \mu_k)^2 p(u) du \quad (3.15)$$

After minimization of distortion, the update equations for μ can be derived as shown below.

$$\mu_k^m = 2t_k^m - \mu_{k-1}^m \quad (3.16)$$

The σ^2 and π for each section are calculated once the optimal mean (μ) assignment has occurred. Iteration stops when the parameters no longer significantly change from iteration to iteration. These estimated values are used as the initial parameter estimates for the EM algorithm. The ALMHQ and EM assume that K is known however, except in controlled studies this is usually not true. The determination of K is termed a cluster validation problem[32] and can be solved using information criteria. The most commonly used information criteria is Akaike information Criteria (AIC). Appendix A describes this approaches along with some examples. Once the parameters have been estimated the quantification portion is complete. The results form the quantification phase are then used as input to the segmentation phase.

The segmentation portion consists of two main steps: maximum likelihood classification (MLC) which performs the initial segmentation, and contextual Bayesian relaxation labeling (CBRL) which performs the final segmentation [26]. The MLC can be used if we treat l_i^* , the true pixel label, as an independent non-random unknown constant. Then the label assignment is performed by maximizing the likelihood for each pixel in the image. The assignment of a pixel i into a class k is given by the following relationships

$$\Gamma(X/\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-(x_i - \mu_k)^2}{2\sigma_k^2}\right) \quad (3.17)$$

$$l_i = \arg \max_k \Gamma(X/\mu_k, \sigma_k^2) \quad (3.18)$$

where Γ is the likelihood function of pixel images for all pixels. The ML estimate of Γ for k would yield estimated k^{th} class label. This is realized by minimizing the log likelihood function given

$$d_{ik} = \log\left(\frac{1}{\sqrt{2\pi\sigma_k^2}}\right) + \frac{(x_i - \mu_k)^2}{2\sigma_k^2} \quad (3.19)$$

where d_{ik} is defined as the Mahalanobis distance between the intensity of pixel i and mean of class k .

$$l_i = \arg \left\{ \min_k d_{ik} \right\} \quad (3.20)$$

Thus, the label of the class mean that is closest to the pixel (in terms of Mahalanobis distance) is selected as the new pixel label.

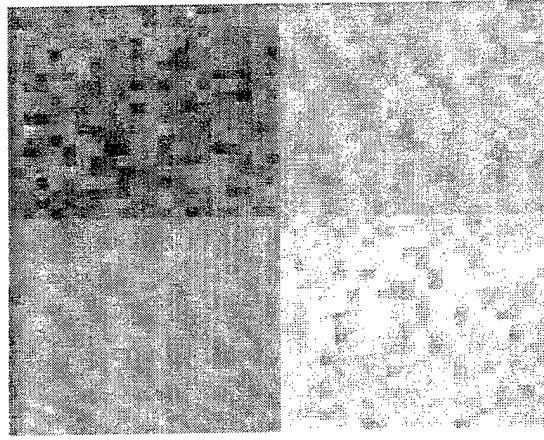


Figure 3.3: Raw four class phantom at 25db SNR

Relaxation labelling methods like CBRL perform efficient segmentation given initial pixel labels. This is accomplished by incorporating contextual information in the segmentation process. Context information is defined as the information relating a label (or class) to a pixel. The contextual information is considered by defining a neighborhood $b \times b$ pixels around the pixel i . The CBRL derivation starts by defining δ_i the pixel neighborhood and l_{δ_i} the labels of the neighborhood. $l_{\Delta i} = l_{j/\delta_i} \quad j = 1, \dots, b^2 \quad j = i$. Next, we can derive the neighborhood membership as

$$\pi_k = \frac{1}{b^2 - 1} \sum_{\Delta i} I(l_i = k, l_{j/\delta_i}) \quad (3.21)$$

where I is the indicator function given by

$$I(x, u) = \begin{cases} 1, & x = u \\ 0, & x \neq u \end{cases} \quad (3.22)$$

π_k can also be interpreted as the conditional probability of l_i . The pdf of the gray level is given by

$$p(x_i/l_{\Delta i}) = \sum_{k=1}^K \pi_k p_k(x_i) \quad (3.23)$$

based on SFNM formulation. The segmentation is performed by minimizing the total classification error using the following relation.

$$l_i = \arg \left\{ \max \left(\sum \frac{I(l_{j/\delta_i}, k)}{b^2 - 1} g(x/\theta_k) \right) \right\} \quad (3.24)$$

where $g(x/\theta_k)$ is the gaussian kernel.

3.3.2 Experimental Simulation

The quantification and segmentation algorithm was simulated with a phantom image and real mammograms. The phantom was a 40×40 image that contained four intensity values (32, 42, 52, 62) each occupying 25% of the image. The image was then corrupted by AWGN that yielded a raw image with a SNR of 25dB as seen in Figure 3.3. The performance of the algorithm was evaluated by the analysis of the quantification and segmentation results. For quantification the true SFNM model parameters were compared to the estimated parameters. These results are depicted in Table 3.2.

From examination of the table the parameters estimates are within 0.5% error for μ and 7.0% error for π . Feeding the parameter estimates into the SFNM model and measuring the GRE between the phantom histogram and model shows that the distribution closely models the image. Finer estimates can be obtained, but the EM algorithm stop criteria must be decreased. In this current arrangement, the threshold is set to 5. By decreasing it to 1, the error

k	μ_k	$\hat{\mu}_k$	π_k	$\hat{\pi}_k$	$\hat{\sigma}_k^2$
1	32	31.82	.25	.242	6.9
2	42	41.79	.25	.2692	9.59
3	52	52.29	.25	.2460	6.7
4	62	62.08	.25	.2429	6.19

Table 3.2: SFNM parameters estimates for four class phantom.

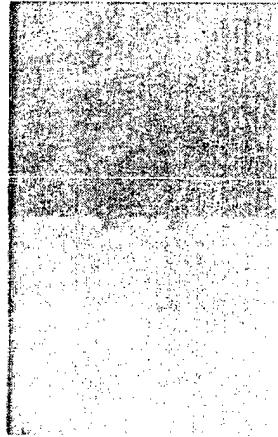


Figure 3.4: Segmented version of four class phantom

percentages drops from 0.5% for μ to 0.3%. This error decease is also accompanied by an increase in processing time.

The results of segmentation of the four class phantom is shown in Figure 3.4. The performance of this portion of the algorithm was judged using the number of pixels in error and the amount of improvement in GRE between the processed and unprocessed images. In this example, the number of pixels in error drops drastically after processing from ____ to _____. This, in turn, improves down stream processing by removing unwanted intensity fluctuations in the image. This segmentation process is not without error. In several simulation runs, it appears that the error pixels are equally distributed across the image with most of the errors occurring between adjacent classes (i.e. pixels from class one are classified as pixel from class two). This appears to be attributed to the resolution of the quantification phase. This is similar to the resolution limitation of a FFT to resolve closely spaced frequencies [77]. If the quantification groups pixels into adjacent classes then the error feeds through into final segmentation.

The mammogram examined was 500×300 with 256 gray levels. From appendix A and [26], the number of classes for typical mammograms are found to be eight. Figure 3.5 shows a raw mammogram and Figure 3.6 shows a segmented version of the mammogram divided into individual classes. Because no ground true tissue map exist for real mammograms the performance will be compared to previous results obtained in [26]. Table 3.3 shows the estimates for the SFNM parameters for Figure 3.5.

These values roughly follow the results presented in [26]. Differences can be attributed to the imaging environment

	1	2	3	4	5	6	7	8
μ	27.39	32.89	62.84	105.17	132.24	159.53	181.45	203.55
σ^2	.46	1.9	294.69	162.75	82.38	81.00	76.04	52.06
π	0.0002	.353	.059	.052	.164	.116	.169	.087

Table 3.3: SFNM parameters estimates for mammogram with 8 classes.

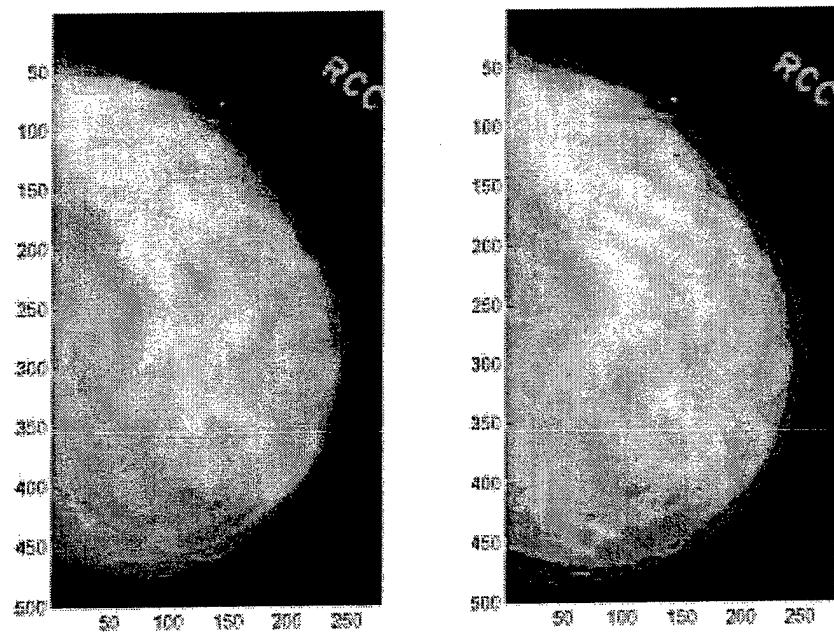


Figure 3.5: Raw and segmented versions of a mammogram

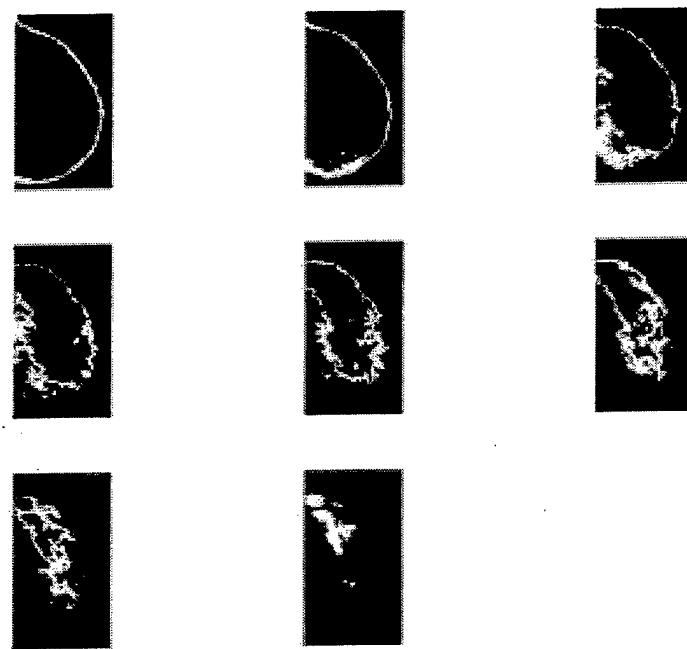


Figure 3.6: Segmented classes from a mammogram

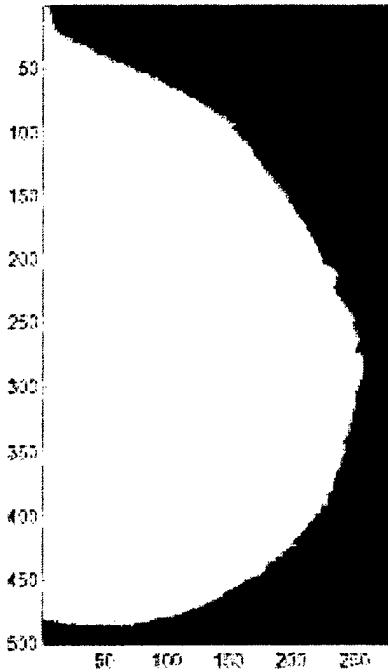


Figure 3.7: Mask image

(i.e. equipment used, signal strength, etc.).

3.3.3 Breast Tissue mask formation

The processing mask is formed by segmenting the raw image into two classes corresponding to tissue and non-tissue (background). Then for every pixel assigned to the tissue class the corresponding pixel location in binary image is set to one otherwise the pixel location is set to zero.

$$Mask(i, j) = \begin{cases} 1, & l_{i,j} = 1 \\ 0, & l_{i,j} = 0 \end{cases} \quad (3.25)$$

This mask image serves two purposes. The first purpose is to limit processing to only tissue regions of the image by multiplying non-tissue pixels by zero. This process increases processing speed and eliminates unwanted background effect in none tissue regions. The second purpose is to feed a morphological filter designed to extract the breast contour for use in further processing. Figure 3.7 shows some typical mammograms with the associated mask.

3.3.4 Contour Construction

The contour is constructed by passing the mask image through two morphological filters. Morphological filters are filters designed through a structuring element to perform different tasks. The structuring element is a $q \times q$ mask where $q \times q$ is smaller than the image size. The first filter is a dilation filter and it has the effect of thickening the object. The second filter is an erosion filter which has the opposite effect (i.e. thinning). The outline can then be formed by subtracting the dilated image by the eroded image yielding the object outline. A flow diagram of this process is shown in 3.8. Figure 3.9 shows some example extracted contours.

3.3.5 Object description

Initially point to point correspondence between images is unknown, but object to object correspondence is known. Using this object correspondence, an initial transform can be derived. Objects in the image include clustered dense

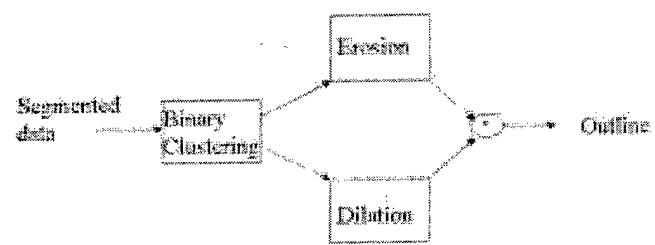


Figure 3.8: Contour extraction process

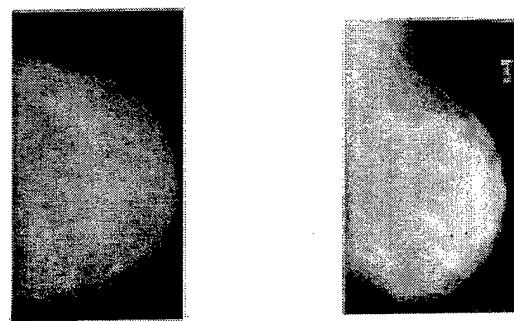


Figure 3.9: Extracted mammogram contours.

tissue and the breast skin line. The first and second moment of the (x, y) coordinates are used to describe the object's geometry. The first moment is calculated using the following equation

$$R_g = \frac{1}{N} \sum_{i=1}^N r_i \quad (3.26)$$

where r_i is the (x, y) coordinate of a single point on the object of N samples and R_g is the center of gravity (first moment) of the object. The second moment is calculated using this relationship

$$C_r = \sum_{i=1}^N r_i r_i^t \quad (3.27)$$

where C_r is the covariance matrix (second moment) of the (x, y) points of the object. To further describe the object, the principle axes and dispersion along these axes is desired. The principle axes of a object is the axes about which the object's entia is minium. The dispersion along the axis is the spread of (x, y) values. The principle axis and dispersion have been shown to describe an object's orientation and scaling [53]. It has also been shown [50] that eigenvalue analysis [86] yields the principle axis and associated dispersions through the eigenvectors and eigenvalues of the covariance matrix of the object. So, the final description contains the center of gravity, principles axes and the dispersion along these axes.

3.3.6 Nipple point estimation

The nipple in most screening mammograms views lies on the extrema of the breast skin line. Several methods exist to determine the extrema point. In [88], the point is estimated by determining the point on the skin line that is farthest from the chest wall line. This method is suspectable to noise in chest wall estimation. Another more stable approach is by [7] which estimates the nipple location through least mean square error approximation of the skin line to a quadratic function. The skin line is obtained using intensity thresholding. The least mean square formulation is shown below.

$$f(x) = c_0 + c_1 x + c_2 x^2, \quad (3.28)$$

$$\begin{aligned} e &= \sum_{i=1}^n (y_i - c_0 - c_1 x_i - c_2 x_i^2)^2, \\ \frac{\delta e}{\delta c_0} &= 0, \quad \frac{\delta e}{\delta c_1} = 0, \quad \frac{\delta e}{\delta c_2} = 0, \end{aligned} \quad (3.29)$$

where c 's are weighting coefficients and n is the number of samples in the contour. The above derivatives yield the following system of equation where c_0, c_1, c_2 are the unknowns.

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - c_0 - c_1 x_i - c_2 x_i^2) &= 0 \\ -2 \sum_{i=1}^n x_i (y_i - c_0 - c_1 x_i - c_2 x_i^2) &= 0 \\ -4 \sum_{i=1}^n x_i^2 (y_i - c_0 - c_1 x_i - c_2 x_i^2) &= 0 \end{aligned} \quad (3.30)$$

This approach is stable for breast skin lines that closely follow the quadratic function which MLO view images generally do not. In this research, the method by [7] is extended by the use of statistical segmentation to extract skin line, and a higher order polynomial as curve fitting function. The nipple estimation procedure is given by the following steps:

- (1) Segment the raw image into classes.
- (2) Group those classes into two classes of breast tissue and background forming a binary image.
- (3) Extract the skin line using morphological filtering.
- (4) Using N contour points $f(x_i)$ of skin line, curve fit a n^{th} order polynomial using least squares. The formulation is as follows:

$$\begin{aligned} f(x) &= c_0 + c_1 x + c_2 x^2 + \dots + c_n x^n \\ e &= \sum_{i=1}^n \left(y_i - \left(c_0 + \sum_{l=1}^n c_l x_i^l \right) \right)^2 \end{aligned} \quad (3.31)$$

Method	<i>x</i>	<i>y</i>
GOOD	289	279
LEHIGH	275	294
WOODS	287	277

Table 3.4: Estimated nipple locations for a CC contour the methods.

Method	<i>x</i>	<i>y</i>
GOOD	345	278
LEHIGH	238	236
WOODS	367	274

Table 3.5: Estimated nipple locations on a MLO contour for the three methods.

This leads to a $n + 1$ system of equation to be solved for the weighting coefficients c .

(5) Find the critical points of $f(x)$ using the following

$$\frac{df(x)}{dx} = 0 \quad (3.32)$$

then solve of x .

A n^{th} order polynomial results in $n - 1$ roots. So, to reduce the number of roots to a manageable number complex roots, zero roots, and roots outside the breast tissue were dropped from analysis. The x yielding the largest $f(x)$ is selected as the skin line extrema or nipple location.

3.3.7 Simulation Experiments

The performance of this algorithm was tested through comparison with the results obtained by [6] and [88]. The skin line contours were extracted using the procedure describe in above section. The algorithms were run on several CC and MLO view mammograms. Table 3.4 shows the x, y location for a representative CC mammogram using the three methods.

Table 3.5 show the x, y location for a representative MLO

In the CC image, our method obtains a nipple estimate closest to the visually selected nipple, but in the MLO image the [88] method selects the best nipple. Our method selects the bottom of the nipple in this case. On average, our method out performs both [6] and [88] because of the low order polynomial used for curve fitting and contour extraction noise. Table 3.6 shows the MSE between a contour and various order polynomials functions for CC and MLO mammograms.

From this we see the higher order functions obtains a lower MSE especially on MLO contour which are not generally quadratic. Thus, with higher order polynomials a more robust nipple estimation is achieved. To further highlight the need for higher order polynomials, Figure 3.10 shows the nipple locations given various order polynomials. The proposed algorithm results were evaluated by radiologists and were found to be accurate in 95 % of the cases. Although some cases estimated the top or bottom of the nipple, the 5 % error can be attributed to contour extraction error. In these cases, the contour was not very smooth causing many local extrema points. This problem could be addressed using a smoothing filter on the contour before nipple detection.

CC	MLO
Order	MSE
2	415.9
5	162.8
10	113.4
20	415.9
Order	MSE
2	3640
5	1419
10	1381
20	1129

Table 3.6: MSE between the contour and n^{th} -order polynomial for CC and MLO views

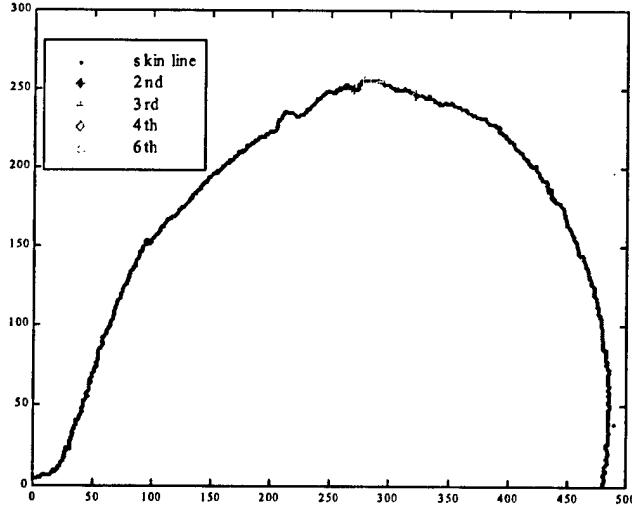


Figure 3.10: Estimated nipple locations for 2,3, 4, and 6th order.

3.4 Site Model Update

In [75] and [13] site model update (or extension) is the process of finding and modeling un-modeled buildings (objects) and adding them to the site database. This is possible because the image-to-site model registration provides the correspondence (overall alignment and camera angle) necessary to compare regions. Once registration is completed, the newly aligned images are then processed looking for set model parameters. These new parameters are compared to existing parameters looking for differences. The differences in parameters are new locations which are then added to the database yielding a composite view of the scene.

In this research, the use of the site model differs from that of [75] and [79] because the site model is used as a reference model with a variable parameter (change map) not a variable model where every parameter could be updated. Site model update, for this application, identifies changes found in new images (registered) and adds them into the site model parameter change map while leaving the other site model parameters untouched. The untouched parameters represent the characteristic of the reference image, and by definition of reference should not be altered. So, overtime this database will contain the reference image information and changes that have taken place over the sequence. This formulation of the site model meets the main objects stated previously which are to provide a common registration frame and highlight the change region for possible exclusion from further processing. Next, the update processes will be explained in more detail.

The site model update process is conducted by modifying the change map (M) parameter with the newly detected change. The change map parameter is an image the same dimension as the scene image where each pixel $M(i, j)$ is initialized to zero to start. Then, each time a pixel $M(i, j)$ is identified as being changed the value of $M(i, j)$ is incremented. Figure 3.11 shows an example change map for a growing object. From Figure 3.11 we see that the object has grown through four images of the sequence. This map could then be used to quantify the change by calculating the size, shape, and rate of change for the object through the sequence.

1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	2	2	2	2	2	1
1	1	1	2	2	2	3	3	3	3	3	1
1	1	1	2	2	2	3	3	3	3	3	1
1	1	1	2	2	2	3	3	3	3	3	1
1	1	1	2	2	2	3	3	3	3	3	1
1	1	1	2	2	2	3	3	3	3	3	1
1	1	1	2	2	2	3	3	3	3	3	1
1	1	1	2	2	2	3	3	3	3	3	1
1	1	1	1	1	1	1	1	1	1	1	1

Figure 3.11: Change map for a 4 image sequence

Chapter 4

Site Model Supported Image Registration

4.1 Introduction

The registration process is supported by the concept of a site model and site model operations. The site model is a mathematical representation of a scene under analysis. A basic site model contains a geometric description of an scenes objects (area, size, and other attributes), raw data, and simple user input (previous tumor locations). The environment interacts with the site model through the site model operations: construction, image-to-site registration and model parameter update. The site model is constructed by thoroughly processing the first image in the sequence to obtain the parameters. The site model supports registration in three main ways. First, the site model forms the reference frame (reference image) for all subsequent images, thus allowing all of the images in the sequence to be alignment to a common coordinate system. Second, the model stores registration parameters like object contours, control points, and user identified regions. This effectively integrates both manual and automatic control objects in a single place. Third, the model stores previously detected change, this enables the current registration process to exclude the previously detected changed portion from the current analysis which improves algorithm robustness. This chapter mostly considers the development of the image-to-site model operation starting with registration theory.

Image registration is the process of overlaying two images with the motivation of transforming one of the images, usually called the float image (I_2), into the same coordinate system as the other image called the reference image (I_1). The process consists of two steps. First, perform a spatial-coordinate transform or mapping function (f) which is used to determine the corresponding coordinate in the new image as shown below.

$$(x', y') = f(x, y) \quad (4.1)$$

In more complex mappings, f can be broken into f_x and f_y corresponding to the x-component and y-components respectively. Typically, (x', y') will not map to an integer grid point on the new image so, some interpolation is need to find the correct (x', y') . The second step of registration is the intensity transform (g), which is used to assign an intensity value to the pixel location (x', y') . Interpolation of the gray levels may also be required to obtain the intensity of point (x', y') . The mathematical expression for registration is given next.

$$I_2(x', y') = g(I_1(f(x, y))) \quad (4.2)$$

Some registration application do not require an intensity transform (i.e. intensity mapping table) such as single modality registration with similar gray level distributions, but multi-modality applications require a more complex transform that accounts for gray level differences between the two modalities.

The key problem in image registration is the determination of the spatial-coordinate transform. The most common types of transforms are rigid (distance between points in the image are preserved under a transform); affine (straight lines and parallelism are preserved between images); projective (straight lines are preserved); and curved (straight line on the original image maps to a curve on the new image). The rigid transform is characterized by a rotation, translation, and scaling which is realized by the following relationship:

$$F = AX + T \quad (4.3)$$

where A is the rotation matrix and T is the translation matrix. This equation can be rewritten as the following

$$f(x, y) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} \quad (4.4)$$

$$a_{11} = a_{22} = \cos(\theta), a_{21} = a_{12} = \sin(\theta), a_{13} = t_x, a_{23} = t_y.$$

The affine transform is more flexible because the a values from the above equation are not restricted to take on only \sin and \cos values. The only constraint is A must be real valued. Projective transforms are realized in a similar manner

$$f(x, y) = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ l \end{bmatrix} \quad (4.5)$$

where w is the extra homogeneous coordinate. Finally, the curved transform is modeled by a n^{th} order polynomial as shown below.

$$f(x, y) = a_{00} + a_{10}x + a_{01}y + \dots$$

In complex mappings, each axis (x-axis, y-axis) has its own polynomial defined as $f_x(x, y)$, $f_y(x, y)$. These polynomials can model several types of transforms. In this research, we focus on the rigid, affine, and polynomial based registration methods to register the sequence of mammograms of the same patient.

Image-to-site model registration is performed by a multi-step algorithm consisting of an initial and final phase. The initial phase registers the images using the principle axis of the skin line in conjunction with segmented internal objects to form a multi-object global spatial-coordinate transform followed by a simple look up table for the intensity transform. The final registration phase consists of a global thin-plate spline transform derived from the control points of the interior breast tissue. The intensity transform in this step is also a look-up table. Next each phase is described in detail, followed by simulation, results, and discussion.

4.1.1 Initial Registration

The main goal of initial registration is to correct for large mis-alignments between images in a sequence. The mis-alignments come from differences in breast placement upon examination, image acquisition process, and film size differences. Although the breast is generally considered a non-rigid object [84], a rigid approach is used as the basis of this phase. This approach is justified by the fact that the distortions, the initial phase is trying to correct, are more or less rigid in nature. In addition, it can be applied without complex knowledge of the input data (i.e. control point correspondence). An example change that is considered in this phase is film size differences. This occurs when different film sizes are used in the acquisition. This type of problem is handled by increasing or decreasing the image by a global scale factor which is addressed by a rigid transform (scaling). The initial registration is performed by a multi-object principle axis registration (PAR) algorithm. The objects include the breast skin line and other extracted internal objects (i.e. clustered dense tissue). The algorithm proceeds as follows: (1) Extract the contours (skin-line and internal objects) from both images. The contours and objects from the reference image are stored in the site model. (2) Use PAR to obtain the transforms for each object. To insure similar objects are extracted from both images, the incoming images are histogram specified to match the reference image (site). (3) Transform each pixel of the image using the transform that is closest in terms of Euclidean distance. This type of transform is called a local rigid transform. The complete process can be summarized into three main steps which are preprocessing, spatial-coordinate transform, and intensity mapping. Figure 4.1 shows a flow chart of the initial registration process. Next, each of these phases are explained.

4.1.2 Preprocess

In this phase, the objects used in initial registration are determined. An object is defined, as a cluster of the same tissue type in the image. Tissue types are identified with statistical segmentation which assigns a label (tissue type) to each pixel of the image [19] [26]. Clusters are identified by using class based region growing where the joining criteria is the pixels class membership. In order to perform registration, some level of correspondence must be established between the images. Visual inspection of extracted objects is used to determine object correspondence. An important step in this process is the identification of similar objects. This problem can become complex when the two images have different pixel intensity ranges. This causes the segmentation algorithm to produce different pixel class assignments resulting in different looking objects. To combat this problem, histogram specification is performed on the incoming image in order to match the site image. In histogram specification, the goal is to adjust

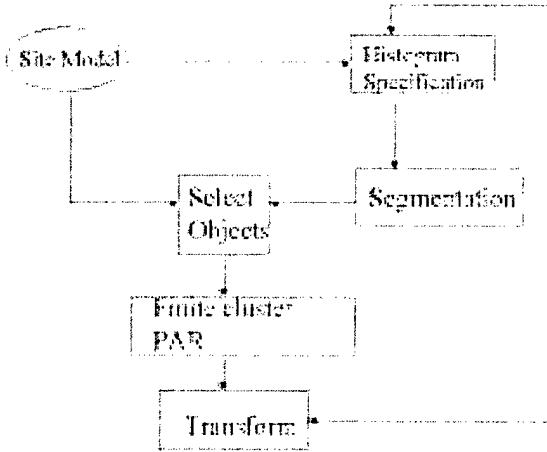


Figure 4.1: Process flow for initial registration phase

the intensities of an image so that the image's histogram matches a desired shape namely the histogram of the site image [85]. The process consists of three steps:

1. Equalize the input image histogram via histogram equalization [85].

In histogram equalization, the raw image intensity values are adjusted to produce a uniform histogram. Consider the pixels x in the image to be random variables with a probability density distribution of $p_x(x)$ and a cumulative distribution of $F_x = P[x \leq x]$. Then an associated uniformly distributed random variable would be $y = \int_0^x p_x(x)dx$. In the digital domain, the integral is replaced by a sum which results in the follow equation.

$$y = \sum_{i=0}^n p_x(i) \text{ where } y \text{ is the new pixel value resulting from the transform } y = T(x).$$

2. Equalize the desired histogram (histogram of site image).

3. Determine the new gray level by matching the pixel value in the equalized image y with the gray level required to make the transform equate to y . $y = G(z) z = G^{-1}(y)$ where z is the new intensity level and G is the transform

Now the histogram specified image is then segmented resulting in more similar looking class assignments.

4.1.3 Simulation Experiments

Next, an object extraction example is consider using the sequence shown in Figure 4.2. This sequence is composed of mammograms of the same patient, acquired at different times. Figure 4.3 shows the class assignment for Figure 4.2. From this figure we see the segmentation did not yield uniform pixel membership across the sequence. Thus, object selection becomes subjective. This fact is highlighted by examining the histograms of the images as shown in Figure 4.4. To correct this problem, the incoming histogram is specified to better match the site image. This is shown in Figure 4.4. This results in a uniform segmentation across the sequence as seen in Figure 4.5. Region growing is then applied to both images to create the objects. Objects from Figure 4.2 are shown in Figure 4.6 and 4.7. The objects are then used in the calculation of the spatial transformation.

4.1.4 Spatial transformation

The transform is calculated by using principle axis methodology[50]. The principle axis method is based on determining and manipulating the principle axes of an object in an image. The principle axis of an object is the axis about which the moments of inertia of the object are minimum. In this method, the objects are registered by matching the principle axes. This approach only works with objects that only vary in rotation and scaling. The rotation factor is represented by the eigenvectors of the data scatter matrix and the scaling factor is address by the ratio of associated eigenvalues of the scatter matrix. Translation is handled by collocating both objects at the origin. The algorithm is as follows: (1) obtain the associated coordinates of the object of interest in both images. (2) Determine the center of gravity object using the following equation.

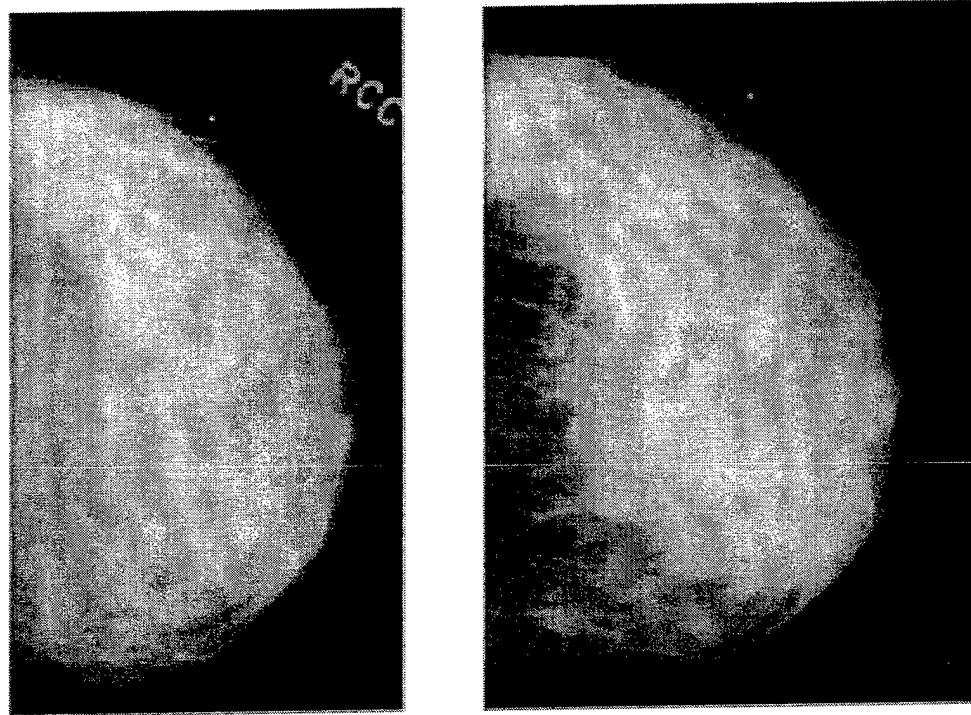


Figure 4.2: Sequence of mammograms

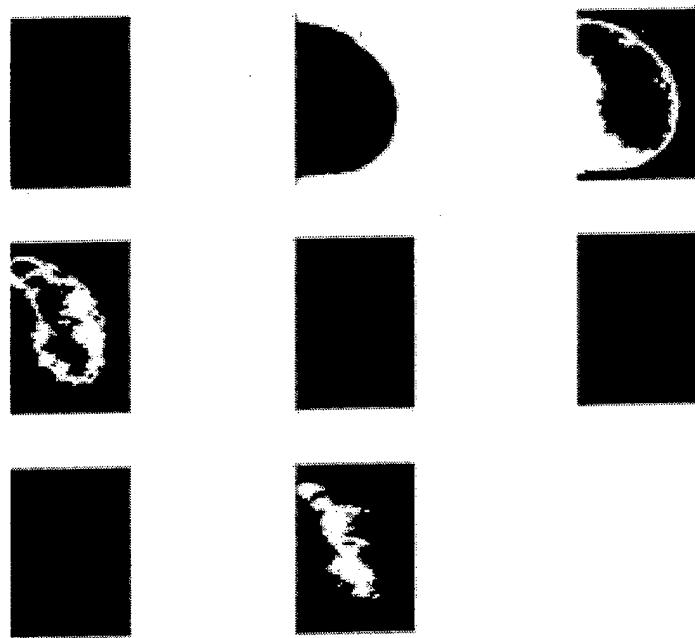


Figure 4.3: Class assignment for raw sequence

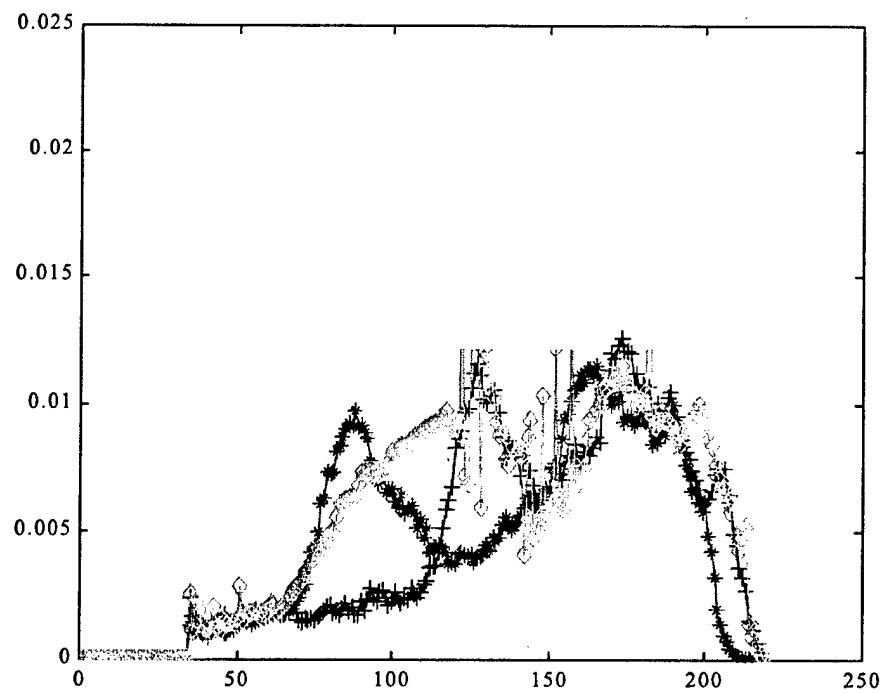


Figure 4.4: Plots of histograms

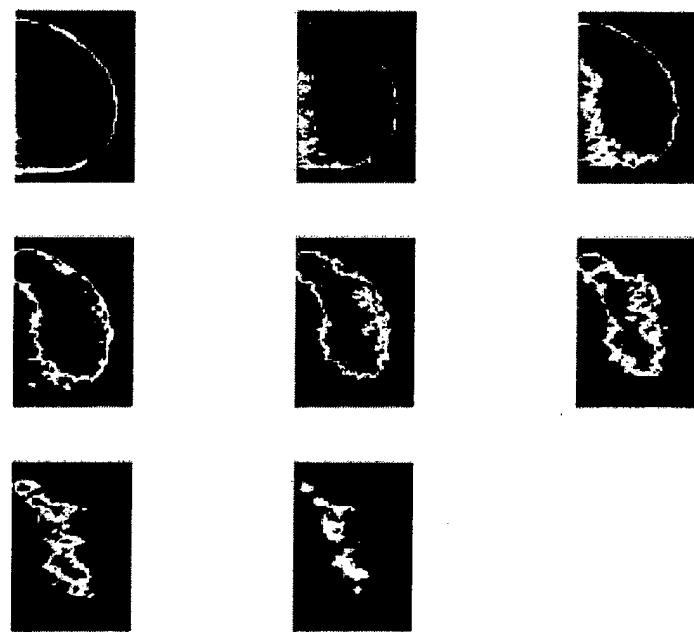


Figure 4.5: Class assignment for specified image

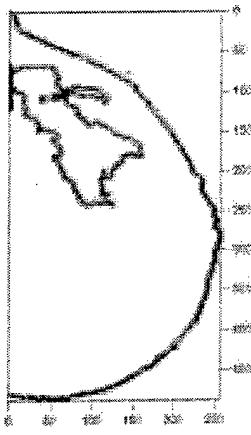


Figure 4.6: Selected object in the site image.

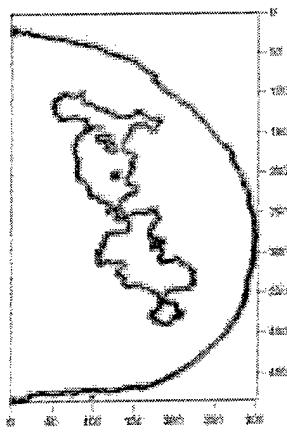


Figure 4.7: Selected object in the float image.

Set	UnRegistered	Registered
1	48.015	37.391
2	45.354	39.613

Table 4.1: Mse between registered and unregistered contours

$$r_{cg} = \frac{1}{N} \sum_{i=1}^N r_i \quad (4.6)$$

r_i represents a point (x, y) and N is the total number of points in the object. (3) Translate the objects so the center of gravity of each object is the origin $(0, 0)$ given by q_i

$$q_i = r_i - r_{cg} \quad (4.7)$$

(4) Calculate the scatter (covariance) matrix of the translated data points q_i 's.

$$M = \frac{1}{N} \sum_{i=1}^N (q_i)^T q_i \quad (4.8)$$

(5) Search for the transformation matrix that diagonalizes M . The transform matrix will be composed of the eigenvectors of M (principle axis). This can be realized by performing singular value decomposition (SVD) of M

$$\Lambda = V^T M V. \quad (4.9)$$

where Λ is a diagonal matrix containing eigenvalues and V contains the associated eigenvectors. (6) Determine the scaling matrix by forming a ratio between the axis dispersion (eigenvalues) of each image.

$$\Phi_f S^2 = \Phi_r \quad (4.10)$$

where Φ is the diagonal matrix containing the eigenvalues and S^2 is a diagonal matrix contain scale factors for each axis. (7) Form the final transform which is a combination of rotation and scaling which is given below.

$$U = V_f^T S V_r \quad (4.11)$$

4.1.5 Simulation experiments

This portion of the system was simulated using the skin line contours of the breast as objects. The derived transform was then applied to the contour points of the float image to obtain a transformed contour. The performance is measured by the MSE between the contours as shown in Table 4.1. Figure 4.8 shows two examples with raw unregistered contours with the associated warped contour. From this table and figure it is apparent that after registration the contours are spatially closer together. The difference between the mse for registered and unregistered is only be about 22%. This is attributed to the end effects where contour points at the beginning and end of the contour create large amounts of matching error. Reducing focus to only consider the central portion of the contour would significantly increase the difference between registered and unregistered mse.

4.1.6 Combination of Spatial Transforms

Assume that multiple corresponding objects can be extracted from the image pair, and from these objects control points could be determined using either contours, surfaces, or object description. In registration, these control points are used to determine a spatial-coordinate transform T that maps pixel in one image to pixel in another. The general expression is shown below

$$x'_i = T(x_i) \quad (4.12)$$

where x'_i is the transformed pixel and x_i is the pixel to be transformed. Three combination approaches have been investigated during the course of this research. Approach one, is a standard approach that considers each of the object pairs as separate registration problems yielding a transform for each object pair. Then a pixel is transformed by a particular transform via some metric Θ (i.e. pixel to contour distance).

$$\begin{aligned} x'_i &= T_k(x_i) \\ k &= \Theta(x_i, T_l()) \quad l = 1, \dots, K \end{aligned} \quad (4.13)$$

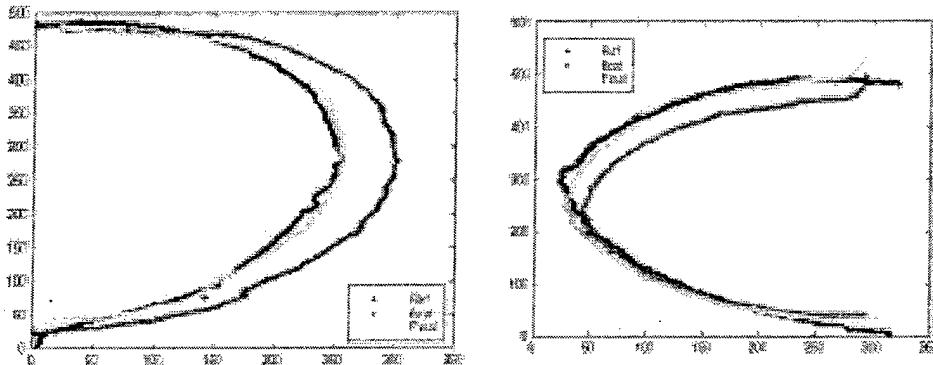


Figure 4.8: Unregistered and registered breast contours.

where k is the transform index ranging from 1 to K number of transforms calculated. This type of transform is called a local rigid/non-rigid transform because pixels are transformed based on transforms local to the pixel [73]. The second approach assumes that each of the T_k describes the same transformation. Then the final transform is obtained by average. The signal model is given below

$$t_i = f_i + w$$

where f is the transform and w is the noise.

Signal averaging is routinely used to improve the signal to noise ratio of signals that are corrupted by noise and can be measured repeatedly [77]. In our case we average the transforms created from all of the objects under analysis to obtain a master transform (T) which is applied to the complete image.

$$T = \frac{1}{K} \sum_{i=1}^K t_i$$

where t_i represents a sample transform and K equals the total number of transforms in the image. This method leads to a global rigid/non-rigid transform since each pixel is transformed by the same matrix.

The third approach, considers the control points as belonging to one of K clusters each with its own mean and variance. Using the mean and variance each cluster can be modeled as a normal distribution. Now, instead of the pixel x_i only being influenced by a single transform it is influenced by a multiple transforms specifically K . The standard transform equation shown above is modified as follows.

$$x'_i = \sum_{k=1}^K \alpha_{ik} T_k(x_i)$$

where α_{ik} is the weighting factor for the i^{th} pixel for the k^{th} transform. This formulation reduces back to the standard transform equation when $\alpha_{ik} = 1; \alpha_{lk} = 0; l \neq i$; Thus each x'_i in this formulation is the weighted sum of K transforms. The weight function could take on several forms such as distance, average, or probability membership. Given that the control points are localized to clusters described by their mean and variance, all of the control point clusters could be made to define a finite normal mixture model as shown below

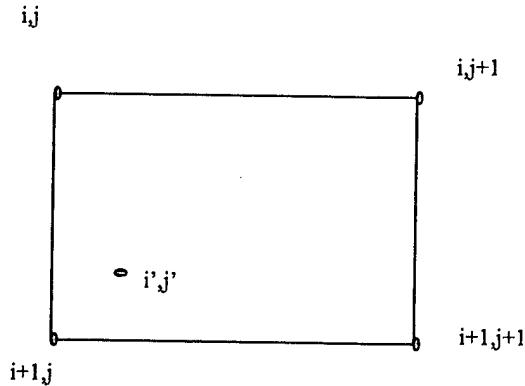


Figure 4.9: Four pixel grid with point (i', j') that falls between the points

$$f(x) = \sum_{k=1}^K \frac{1}{K} g(x/\mu_k, C_k)$$

where g is a gaussian kernel and μ_k and C_k are the class mean and covariance respectively. The mixture model sets the framework for using pixel membership as a weighting criteria. Membership in this context is defined as which transform is used to transform a pixel. This model has been used in image segmentation to determine pixel class [19] [28] [54]. Similar to [19] [28] [54] the posterior probability is used as a measure of each pixels probability membership. The statistical membership of a pixel with respect to a control point cluster can be defined as

$$\alpha_{ik} = P(T_k/x_i) = \frac{g(x_i/\mu_k, C_k)}{\sum_{l=1}^K g(x_i/\mu_l, C_l)}.$$

Thus each pixel in the float image can now be transformed using a membership weighted transform. The gray levels of each pixel are assigned using a straight look up table. The procedure is the following: (1) transform the pixels located at a point (x, y) in the reference image (R_I) to a point (u, v) in the float image (R_f) using the selected transform (T).

$$(u, v) = T(x, y)$$

Determine the intensity at point (u, v) . Since points (u, v) are generally not integer values (i.e. fall on a grid point), interpolation is required to select the intensity. Figure 4.9 highlights an example which requires interpolation. Several interpolation method exist, but for this research Nearest Neighbor interpolation is used. This method assigns the new value (u, v) from the closets grid point surrounding it. This leads to the following relationship.

$$w(x, y) = R_f(T(x, y))$$

4.1.7 Simulation Experiments

The implementation of the following methods are discuss through some examples. Figure 4.10 shows the original image pair under consideration. The image pair was created by the addition of a Gaussian filtered block and rigidly rotating the complete image by 10° . This is a small rotation, but should highlight the effect of the local and global multiple object transforms on the image. Figure 4.11 and 4.12 shows the resulting image pairs after transformation by the local rigid and global rigid transform respectively. From examination of Figure 4.11 it is apparent that discontinuity resulted from the transform as seen on the left hand side of the right image in Figure 4.11. These discontinuity can be attributed to differences in transform used on adjacent pixels. The global registration pair, on the other hand, has a smooth look because of the use of a single transform. So, no more cases of adjacent pixels

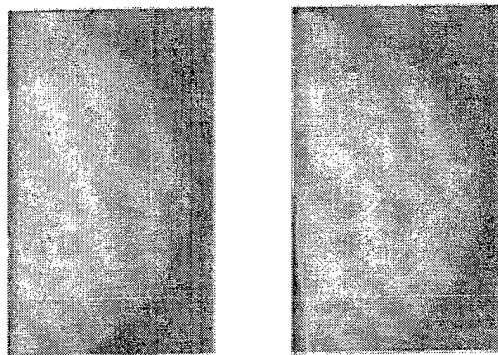


Figure 4.10: Original image pair

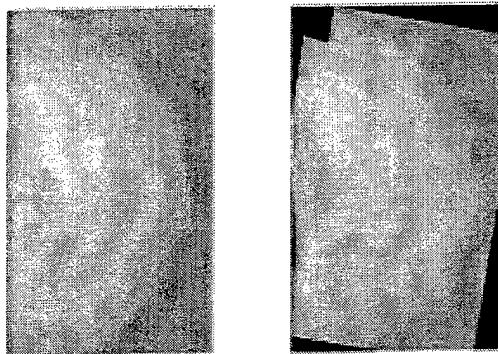


Figure 4.11: Image pair transformed using local rigid with three objects

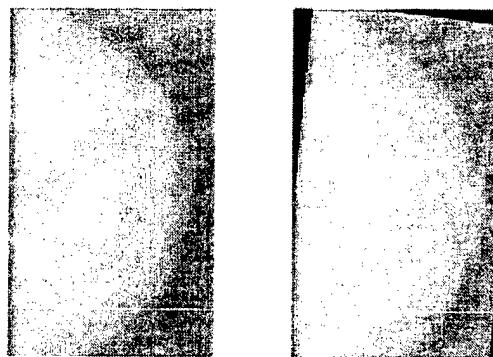


Figure 4.12: Image pair produced with the global rigid

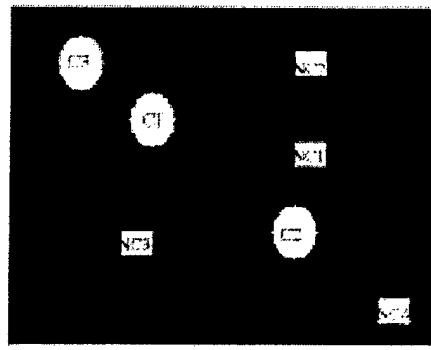


Figure 4.13: Phantom image used in finite normal mixture registration

Object	Configuration		
	1	2	3
C 1	5	7	20
C 2	100	4	5
C 3	20	25	10
NC 1	10	13	4
NC 2	20	6	24
NC 3	5	20	11
NC 4	5	20	11

Table 4.2: Angle rotations for each object in phantom registration image

Number of Objects	Config. 1	Config. 2	Config. 3
0	1616	1598	1785
1	960	930	468
2	758	842	410
3	279	504	310

Table 4.3: Mse results for each configuration

being transformed by different transforms. To simulate the finite mixture registration method, we considered a 150 x 150 phantom image containing three control objects and four non-control objects as seen in Figure 4.13. The control objects are ellipse while the non-control objects are squares 10×10 . The key thing about the control objects is that only object correspondence is known not point correspondence. Each of the control and non-control objects are rotated and translated by different amounts. This simulates a non-linear deformation (non-rigid) between image sets, and serves to test the combination ability of this registration method. The objects rotation angles are given in Table 4.2.

Three configurations of rotation angles are considered. These configurations are chosen arbitrary to show the robustness of the proposed algorithm. In each configuration the image is registered using one, two, or three transforms. The performance is measured in mean square error (mse) between the reference and warped image where a lower mse is seen as better performance. Table 4.3 shows the mse for each configuration. From the table it is apparent that registration by one transform on average reduces the mse by 50%. The mse is decreased another 10% with the addition of another transform. With the addition of the last transform, significant improvement in mse is achieved. The mse is reduced by approximately 75%. Figure 4.14 shows an example of the reference and warped image using all three transforms. These results show the benefit of using multiple transforms where possible.

4.1.8 Final Registration

The goal of this section is to fine tune the alignment achieved in the initial phase by considering the breast as a non-rigid body. This allows for the consideration of the deformation between the image and site model. Deformations are caused by positioning differences subject weight gain, natural growth, and nonuniform compression during examination. To handle these deformations, more complex transforms are required. In [68], the polynomial based transform were shown to be able to handle non-rigid deformation of kidneys so they are selected in this study to model the deformations of the breast. Various types of polynomial transforms exist such as linear, quadratic, and cubic [68]. In this research, a thin-plate spline polynomial will be used as the mapping function [5].

The key requirement for use of polynomial based transform is the existence of control points. In some environments control points are easily obtained (brain images), but in mammograms this task is very difficult because of lack of anatomical landmarks between images. In this research, the cross points between vertical and horizontal elongated structures are used as potential control points. These elongated structures represent blood vessels and milk ducts. To use these points, one must assume they are time and shift invariant for the most part. These points will be defined as potential control points. Then the potential control points are matched to produce the final control points which are then used to calculate the thin-plate spline polynomial transform. The fine registration process concludes with the transformation of the complete image pixel by pixel.

Similar to the initial registration , final registration can be divided into several parts. They are preprocessing, point correspondence, spatial coordinate transform, and intensity mapping. Figure 4.15 shows the complete

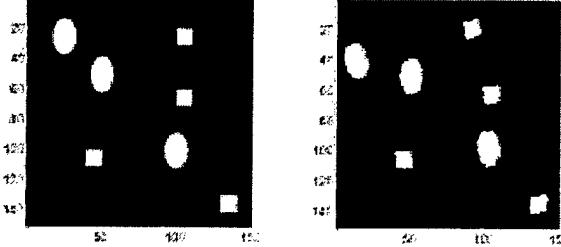


Figure 4.14: Reference and warped image from multi-object registration

algorithm flow. Next, each part will be discussed in detail.

4.1.9 Preprocessing

In this part, the potential control points are extracted from the image. This is achieved by detect the elongated structures in the image using modified monotony operators to highlight both horizontal and vertical structures in the image[7]. The monotony operators are defined by two overlapping rectangular neighborhoods, one small and one large, centered around a pixel (i, j) . Figure 4.16 shows an example of both the vertical and horizontal operators in a image. The operators work as follows: the pixel at (i, j) is labeled one if the number of pixels in the large neighborhood that are larger than g_{\max} , exceeds a threshold τ . Otherwise, the operator assigns a zero to the pixel (i, j) . g_{\max} is defined as the maximum gray level in the small neighborhood surrounding the pixel (i, j) . The vertical and horizontal operators are defined by the following relations

vertical:

$$\begin{aligned} a &= \{(k, l) | k = 1, -p \leq l \leq p\} \\ A &= \{(m, n) | m = 1, -q \leq n \leq q\} \end{aligned} \quad (4.14)$$

horizontal:

$$\begin{aligned} a &= \{(k, l) | l = 1, -p \leq k \leq p\} \\ A &= \{(m, n) | n = 1, -q \leq m \leq q\} \end{aligned} \quad (4.15)$$

$$q > p, \tau = (q - p) \quad (4.16)$$

where a is the small neighborhood of length p and A is the large neighborhood of length q . Using the vertical and horizontal binary images the potential control points are obtained by finding the cross points of vertical and horizontal elongated structures. This is implemented by applying a logical AND operation to the vertical elongated structures image Λ and horizontal elongated structures image Γ yielding Υ image which only contain cross points.

$$\Upsilon = \Gamma \odot \Lambda \quad (4.17)$$

Depending on elongates structure thickness the cross points could contain multiple pixels. In cases like these, the centroid of the group of pixels is defined as the potential control point.

Following the method defined in [7], a Gaussian kernel is passed over the image several times to blur the image in an effort to reduce the effects of fine details in structure detection. This leads to detection of only the most prominent elongated structures. Applying this process to raw images produces an intractable amount of potential control points[7]. Figure 4.17 and 4.18 shows a raw and blur image with their respective elongated structure images.

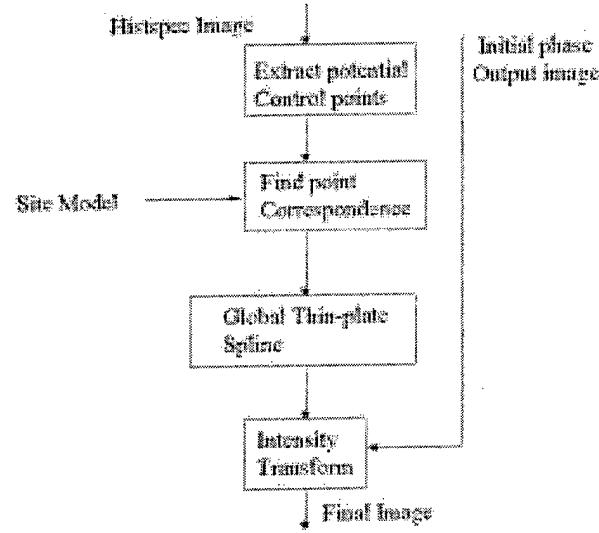


Figure 4.15: Process flow for final registration phase

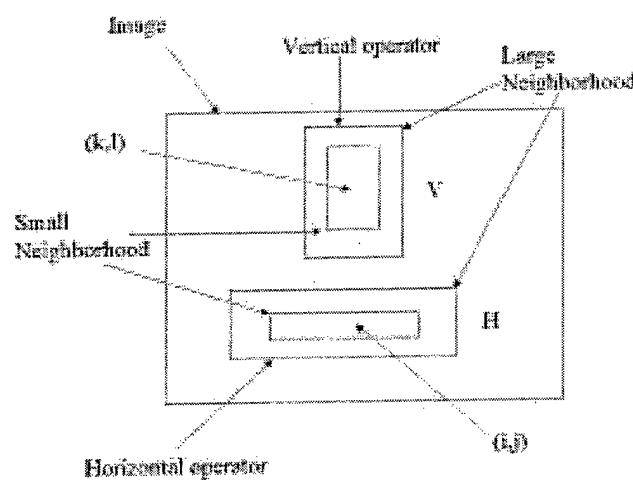


Figure 4.16: Monotony operators for an image

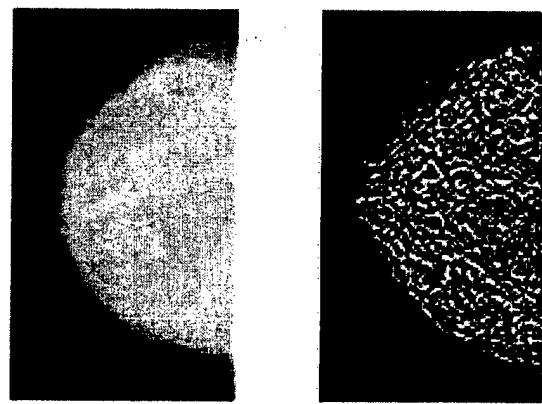


Figure 4.17: Raw mammogram and associated elongated structures

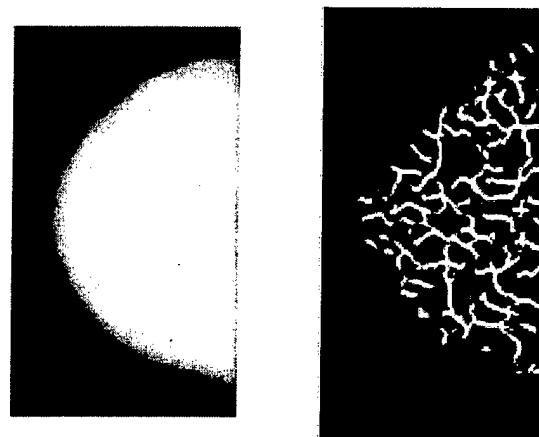


Figure 4.18: Three pass filter mammogram with associated elongated structures

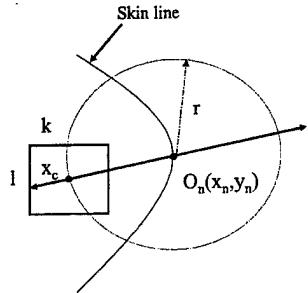


Figure 4.19: Matching window location on new mammogram

4.1.10 Point Correspondence

The next step in the fine registration process is matching corresponding control points from the associated pools of control points in each image. Several methods for point correspondence have been investigated and proposed by [7]. These included a signature matching, which is an algorithm that search for longest direction of an elongated structure cross point, and a wavelet based approach that examined localized regions. In addition, [5] used laws texture features to determine correspondence. This research presents two new correspondences methods. The first is based on the signature matching algorithm by [7], but an attempt is made to match the complete structure not only longest direction. The second method transposes the new potential control points $O_q(x_q, y_q)$ onto the old image and matches control points based on point distance from an old potential control point $O_p(x_p, y_p)$. To improve matching rates on both methods, only a subset of the potential control pool from the new image are tested at a single time. This subset is identified as potential control points contained in a $k \times l$ window centered around the point X_c .

The point X_c is the intersection point between a circle centered around the estimated nipple location $O_n(x_n, y_n)$ in the new image and a straight line passing through O_n with a slope of m as shown in below. The slope m of the line is equal to the slope of a similar line in between the potential control point $O_p(x_p, y_p)$ in the site model (old image) and O_o the nipple location in the old image.

$$y = m(x - x_n) + y_n \quad (4.18)$$

$$m = \frac{y_p - y_o}{x_p - x_o}$$

$$(x - x_n)^2 + (y - y_n)^2 = (x_o - x_p)^2 + (y_p - y_o)^2$$

Figure 4.19 shows a pictorial example. Next, each correspondence method will be discussed.

4.1.11 Elongated structure matching

After passing the location criteria ($k \times l$ window), signatures for each potential control point contained, in the local window, are calculated. The signatures are designed to capture the characteristics of the elongated structures surrounding a potential control point. The signatures are calculated by forming the elongated structure image which contains both vertical and horizontal structures. This is realized as a logical OR operations on the vertical and horizontal structure images as shown below.

$$\Omega = \Gamma \oplus \Lambda \quad (4.19)$$

The image Ω now contains cross points and associated vertical and horizontal elongated structures. Figure 4.20 shows some elongated structures derived from a mammogram.

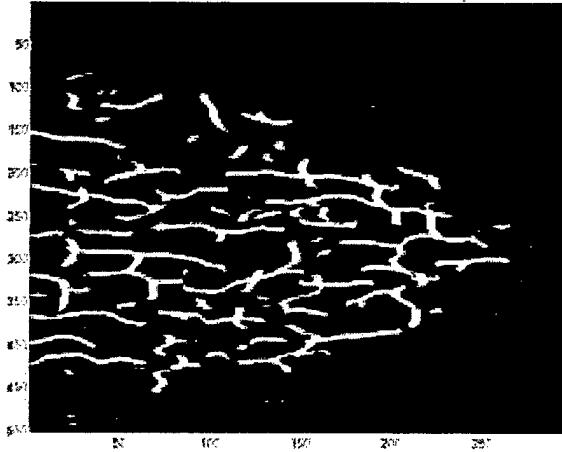


Figure 4.20: Elongated structures detected by monotony operators

The next phase of signature construction is the rotation of a $m \times n$ window N_s steps around the control point. This yields $\Delta\xi^\circ$ for each step. For each step the number of nonzero pixels (NZ) contained within the sum window are counted. The number counted for each step is the signature $y(\Delta\xi^\circ) = NZ$. This process is shown in Figure 4.21. The signatures are then matched by measuring the Pearson correlation coefficient [14] between a pair of potential control point signatures. The resulting coefficient is then applied to a threshold. The Pearson correlation coefficient is formulated by the follow equations

$$\rho = \frac{SS_{xy}}{\sqrt[3]{SS_{xx}SS_{yy}}} \quad (4.20)$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{N_s}$$

where y is the N_s point signature of O_p . The Pearson coefficient measures the statistical distance of two distributions. Because non-rigid deformation occurs between images the corresponding control point signature could be a circularly shifted version of each other as seen in Figure 4.22. To consider this problem, the complete signature of the new image control point is circularly shifted by one sample and then Pearson matched. The highest Pearson between all shifts is taken to be the resulting Pearson value for that (O_p, O_q) pair.

The Pearson results for a (O_p, O_q) pair are stored in a modified accumulator matrix. The accumulator matrix is a $N_o \times N_n$ matrix where N_o and N_n are the number of potential control points in the site model (old) and new images respectively. In traditional accumulator formulations [7] ??, the element (O_p, O_q) is incremented each time point O_p matches point O_q , but in this research we put the maximum Pearson correlation coefficient the element corresponding to (O_p, O_q) . The final match is performed by taking the maximum value down the columns and zeroing the other column entries for that column. This is followed by taking the maximum value in each row and zeroing the other row entries. The resulting matrix should contain only one nonzero value per row and column. The nonzero elements are the control points.

4.1.12 Simulation experiments

Pearson based control point matches were obtained for the phantom and several real mammograms. The phantom sequence was composed of two versions of the same image. The second image in the sequence was a rigidly transformed copy of the first image. The real sequence contained two images of the same patient acquired at different times. Figure 4.23 shows the potential 'o' and real control points '*' for the phantom sequence where 37 out of the 43 potential control points where matched across the sequence. Compare this to Figure 4.24 where only 5 out of the 36 potential control points where matched. This difference in final control point matching is the result of the variability of extracting elongated structures from mammograms. In Figure 4.23, the structures remain

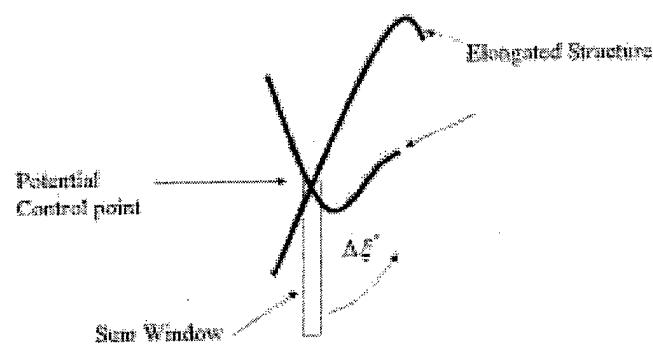


Figure 4.21: Formation of potential control point signature.

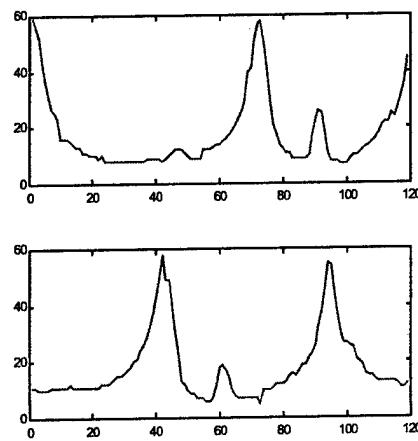


Figure 4.22: Potential control point signature with corresponding shifted version

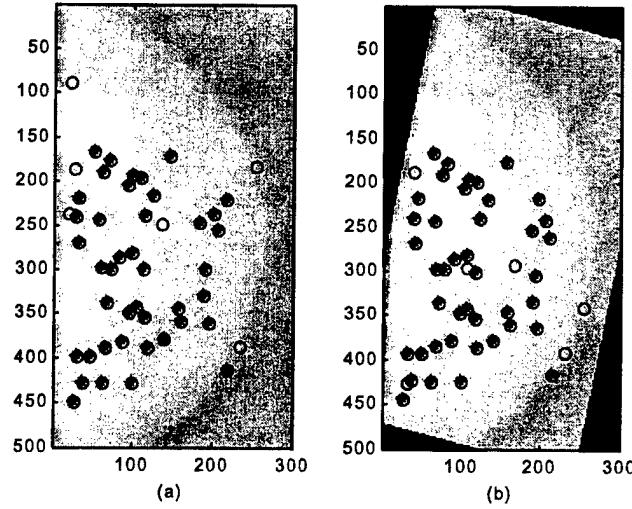


Figure 4.23: Potential and Matched Control points via Pearson matching for the phantom study

stable because the rigid transform causes the signatures to be rotated versions of each other which allows for easy matching. But in Figure 4.24 non-rigid deformation between the image causes the signatures of potential control points to look drastically different if detected at all. In [58], which uses much the same approach but only considers a 40×40 window using the longest arm of the structure as the matching metric, only obtains 6 control points for a real sequence. In this research, a smaller 10×10 window is used along with the Pearson matching criteria to obtain comparable results. This reduction in window size is attributed to use of the complete signature information in matching not just the most dominate structure arm. To increase matches, the local match window currently at 10×10 should be increased. It should be noted that this operation also increases false match probability and processing time.

4.1.13 Nearest Neighbor match

In this method, initial registration is assumed to have corrected most of the global distortion and mis-adjustment between the two images. The control point correspondence is then obtained by overlaying the potential control points from the new image with the potential control points of the old image and calculating the Euclidean distance from each old potential control point to each new potential control point.

$$d_j = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

with i and j equal to the index of potential control points bounded by $i = 1.....N_{old}$ and $j = 1.....N_{new}$. The new potential control point with the smallest d value is selected as a match for the old point of interest. Figure 4.25 shows a typical case of a localized window. In the event, a new potential control point is matched to several old points the match with the smallest d is selected as the final match.

4.1.14 Simulation Experiments

Figure 4.26 shows the same sequence shown in Figure 4.24 where nearest neighbor matching is used. This matching methodology more than doubles the number of matched control points over matching with Pearson matching method. It also produces control points that are distributed evenly around the image. This method exceeds the method presented by [7] at smaller matching window sizes. A key note is the dependence of this method on initial registration. Without initial registration, distance is not a good enough metric along. Again more matches can be obtained by increasing window size at a cost to processing time and false match rate.

4.1.15 Spatial-coordinate

The main goal in registration is to obtain a transform T_A such that one of the images could be transformed into correspondence with the other. In general, an image mapping transform is represented by

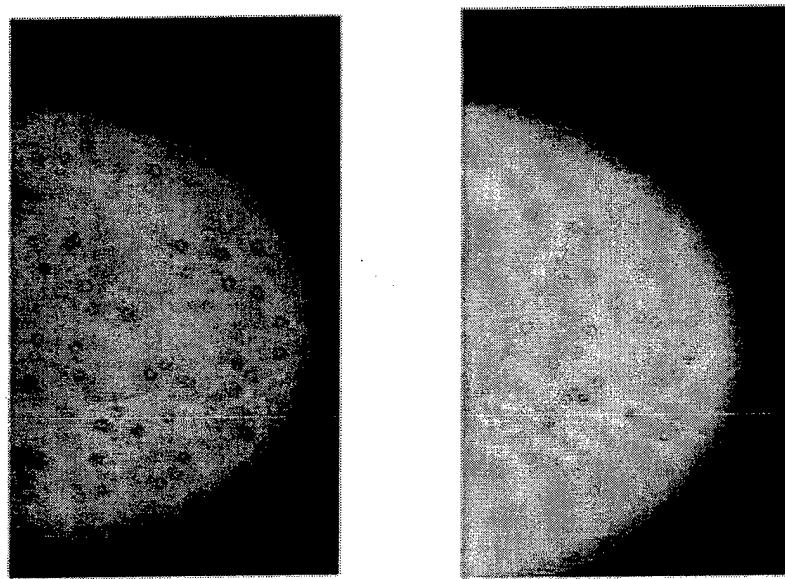


Figure 4.24: Potential control points shown by \circ and matched control points shown by $*$ via Pearson matching

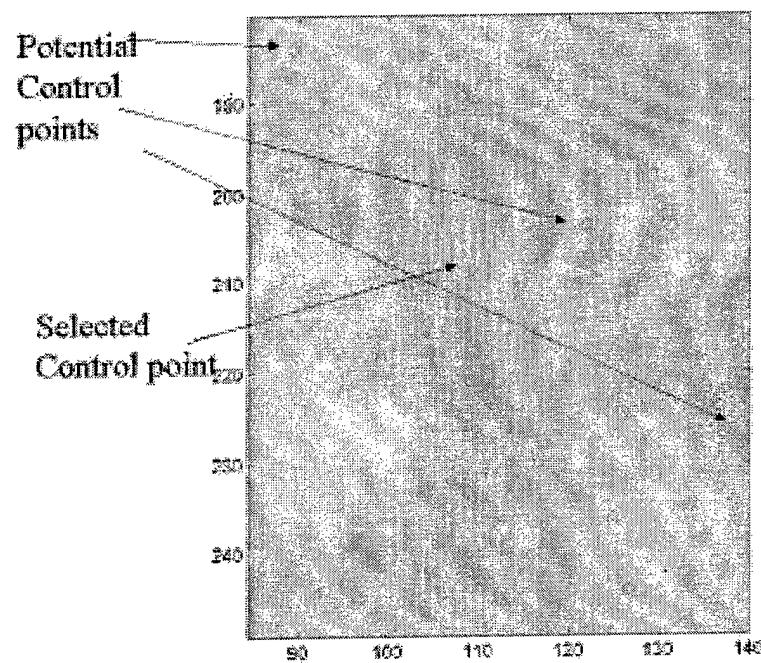


Figure 4.25: Local correspondence window for a potential control point

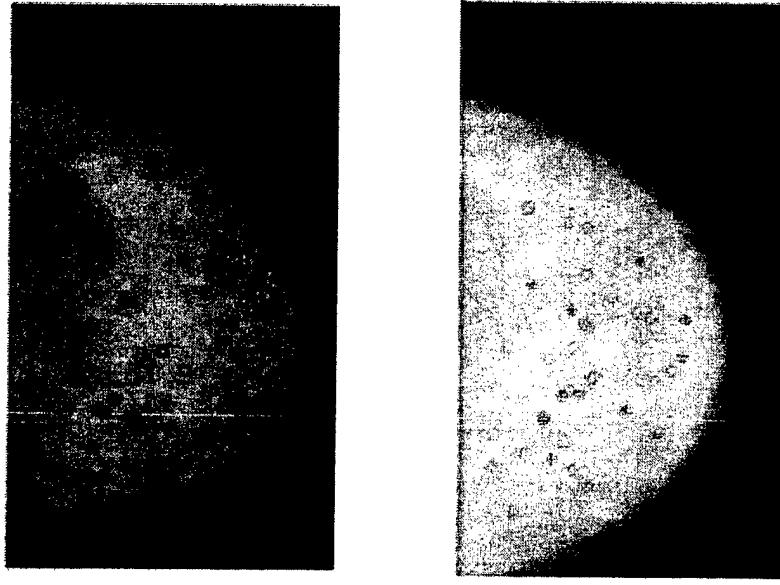


Figure 4.26: Potential control points shown by o and matched control points shown by $*$ via Nearest neighborhood method.

$$T_A(x, y) = (f_x(x, y), f_y(x, y)) \quad (4.21)$$

where $f_x(x, y)$ is the mapping function for x coordinate of (x, y) and $f_y(x, y)$ is the mapping function for the y component of (x, y) . Since breast tissue is inherently nonrigid, complex changes can occur between the image in the sequence. To account for these changes, the function $f()$ needs to be non-linear. [5],[68] selected TPS as the mapping transform so we apply it in our case. The mapping function for TPS is shown below

$$f(x, y) = w_0 + w_1 x + w_2 y + \sum_{i=1}^n W_i g(r_i) \quad (4.22)$$

$$g(r_i) = r_i^2 \log r_i^2$$

given that $r_i = (x_i - x)^2 + (y_i - y)^2$. This transform is made up of a global (affine) portion and (elastic) portion. These two portions are distinct but can be evaluated simultaneously.

In order to use $f(x, y)$ to transform the image, the coefficients w_0, w_1, w_2, W_i must be estimated. This is done by using the control points determined from the previous section, to formulate a least square approach to coefficients estimation. The least squares formulation starts with coordinate mapping relation

$$(u, v) = (f_x(x, y), f_y(x, y)) \quad (4.23)$$

where (u, v) is a point in the new image (control point) that is associated with the point (x, y) in the old image (control point). Given (u, v) and (x, y) are control points, zero error should occur when transforming (x, y) through the mapping function.

$$(u, v) - (f_x(x, y), f_y(x, y)) = 0$$

Rearranging terms and expanding to handle n control points a general error equation is formed given below.

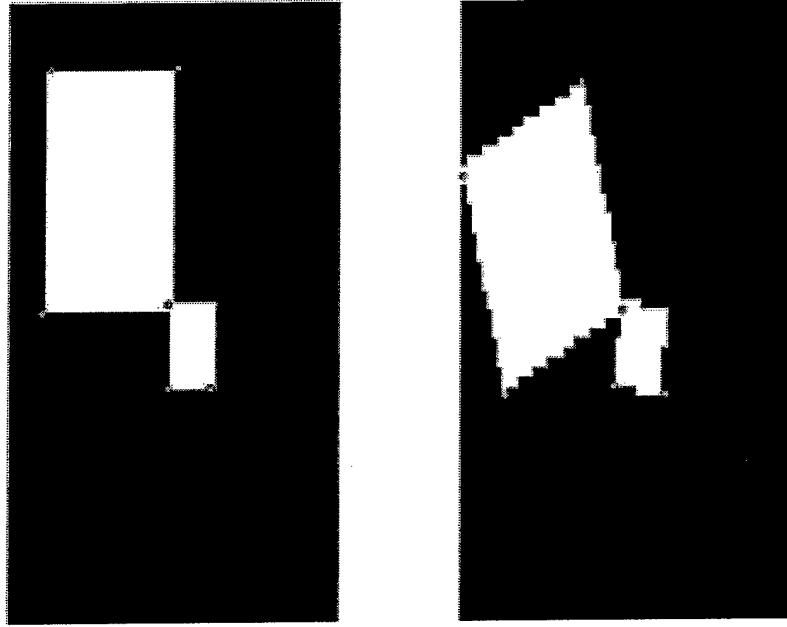


Figure 4.27: Raw phantom sequence

$$E = \sum_{i=1}^n [(u_i - f_x(x, y))^2 + (v_i - f_y(x, y))^2] \quad (4.24)$$

The above equations leads to the normal equations. The relation for the x mapping functions is shown below.

$$\sum_{i=0}^m \sum_{j=0}^i a_{ij} \left[\sum_k x_k^j y_k^{i-j} x_k^\beta y_k^{\alpha-\beta} \right] = \sum_{k=1}^n u_k x_k^\beta y_k^{\alpha-\beta} \quad (4.25)$$

where $\alpha = 0 \dots m$ and $\beta = 0 \dots \alpha$. The coefficients for the y mapping functions are found in a similar fashion. With the mapping functions f_x and f_y each pixel is then transformed to produce the warped image. In general, the new pixel location will not fall on a exact grid point some interpolation is used to obtain the pixel value. In this research, nearest neighborhood interpolation is used to determine the new pixel value.

4.1.16 Simulation experiments

This process is examined through the following example of a phantom that is made up of two squares where each square is transformed by a different amount. The image pair is shown in Figure 4.27. Table 4.4 shows the mse between the reference and the stages of the warped image. From the table one can see the mse decrease through out the process. Use of PAR along reduces the mse by 77%. With the addition of TPS the mse is reduces by another 10%. A small decrease in mse after PAR is attributed to the use of only 6 control points. If more control points had been selected the performance gain of TPS in this process should improve.

4.2 Summary

This registration approach is composed of two main steps an initial step and fine step that are supported by the site model. The site model supports the registration process by storing user (manual) and automatically extracted

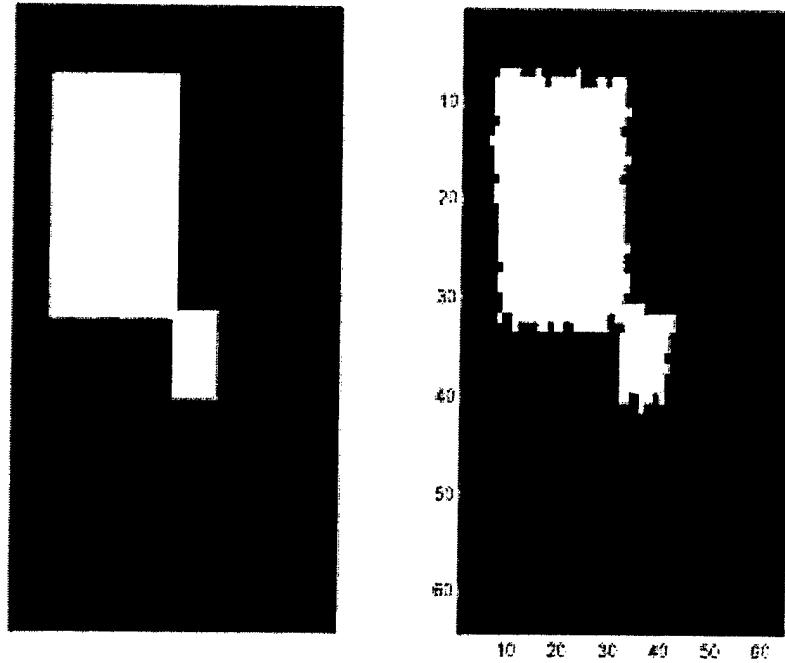


Figure 4.28: Registered phantom sequence

Number in error	Method
395	none
90	PAR
60	PAR-TPS

Table 4.4: Amount of pixels in error for registrations methods

Chapter 5

Site Model Supported Change Detection

5.1 Introduction

Change detection is the process of identifying significant differences as measured by a metric between two or more objects. In this research, the objects of interest are images or sub-images (i.e. localized windows) in a sequence. In an image sequence with objects, three types of change can be defined. In the first type of change, defined as type I, only intensities of the pixels change. In the second type, defined as type II, the intensities remain constant, but the location or shape of the object changes. In the third type of change, defined as type III, intensities, shape, and location change. These types of change can be measured either pixel by pixel or image by image. A simple formulation of a pixel change metric is shown below.

$$D = \mathfrak{S}(R_f, R_r) \quad (5.1)$$
$$\text{image}(i, j) = \begin{cases} 1, & D(i, j) \propto \gamma \\ 0, & D(i, j) \propto \gamma \end{cases}$$

where D is a change map containing the metric measurements at each pixel. $\mathfrak{S}()$ is the pixel function criteria applied for processing. For example, in difference analysis the function \mathfrak{S} would equal abs . γ is the metric threshold, R_f is the transformed image, and R_r is the reference model image. Image change is measured in much the same way as pixel, but the image is evaluated as a whole.

$$D = \mathfrak{S}_O(R_f, R_r) \quad (5.2)$$
$$\text{image} = \begin{cases} 1, & D \succ \gamma \\ 0, & D \prec \gamma \end{cases}$$

where \mathfrak{S}_O is the image change function, D is a scalar change value, R_f is the float image, and R_r is the reference image. An example of an image change function could be the mutual information between to image blocks as shown below where

$$\mathfrak{S}_O = \sum p_{xy} \log p_{xy}$$

p_{xy} is the joint distribution of an image x with marginal density p_x and an image y with a marginal density of p_y .

Change detection in images has found application in various fields including video sequence processing; satellite imaging; and medical imaging. In video sequence processing, numerous change detection metrics have been developed [10]. The main goal in this application is to find abrupt scene changes to aid in sequence compression. The compression is achieved by sending only a reference image (i.e. first image in sequence) then only scene change information (global) in subsequent transmissions. The video change metrics assumes high SNR and the occurrence of abrupt change. The main motivation is to detect the region of the image that contains the change. No effort has been put into describing the change. The most research on change detection has been conducted in the satellite imaging area (remote sensing). In this area, work has been done on building change detection, agriculture crop analysis, and weather tracking [79], [82]. Some specific change metrics have been developed for synthetic aperture (SA) images [83], but they take advantage of the multi-spectral data that is inherent to SA imaging. For this reason, they are not as useful for other applications (i.e. non-SAR applications). Again, as in video change, no effort has been put into describing the type of change.

In the medical environment, the existence of change and the classification of change are very important. This change leads to valuable diagnosis information. Since the change metrics for video requires high SNR and the metrics for SRA are SAR signal dependent, a new metric is needed. The newly developed change process should also have

input for use in registration. The model also provides a common frame for incoming images to register to. Finally, the site model stores the complete image sequence history in a common place. The initial registration step is aimed at addressing gross misalignment between the images. This step is rigid model deformation based and requires little environment knowledge (i.e. control point locations). While the fine registration step requires the identification of corresponding control points. The fine step is aimed at correcting non-rigid deformation between images in the sequence. Together mammograms can be robustly registered in support of change analysis. With the mapping functions derived above each pixel is transformed to produce the new image.

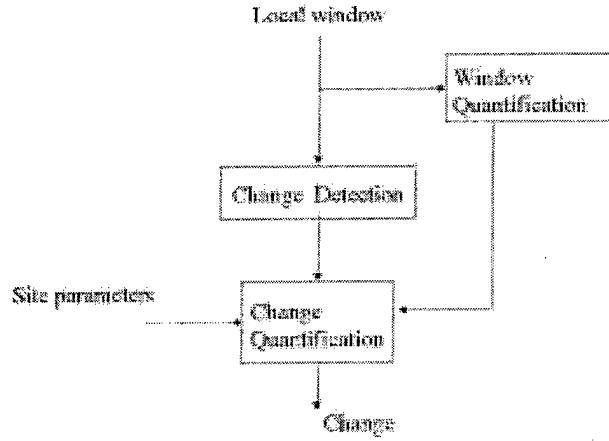


Figure 5.1: Change detection process flow

the capability to quantify change. To accomplish this task, a two step process is developed. The steps are change detection and quantification. The detection phase is performed by measuring the joint relative entropy between the two objects with entropy values higher than a user specified threshold marked as change. Quantification consists of comparing the objects area and center of gravity. Often, in medical applications such as lesions monitoring, change overtime is of great interest as it can show response to drugs or disease progression. The site model, which is a dynamic mathematical and geometrical description of a scene under analysis, has been shown to be a useful tool in the analysis of changing images in a sequence [79]. When applied to the medical change problem, the site model could store the behavior of an object in the scene as well as over all image behavior. Use of the site model also allows the integration of user supplied input (domain knowledge) with automatically extracted parameters towards the goal of change detection. This idea of user input models the real process that a radiologist uses to analysis the image. Specifically, the site model supports change detection in two main ways. First, the site model provides a unified location to store change that has been occurring overtime. This feature is useful in monitoring application. Second, the site model can be used to determine which part of the image should be considered in processing. For example, the changed portion of a image might not be included in transform calculation. This generates a more robust transform. The rest of the chapter considers the development of the change algorithm. The complete block diagram is shown in Figure 5.1.

5.2 Change Analysis Theory

The site model supported change detection algorithm contains two main phases. Phase I is change detection and phase II is change quantification. Change detection is the process of determining whether two objects (images) differ. In practice, nothing is ever exactly the same, so the change detection results are measured in comparison to a threshold. For this research we selected the use of relative entropy as our change metric. Relative entropy is a measure of the inefficiency of assuming that one distribution exactly matches the other. (i.e. distance apart) Relative entropy is given by the equation.

$$D(p//q) = \sum p(x) \log \frac{p(x)}{q(x)}$$

where $p(x)$ and $q(x)$ are the distribution of image P and Q respectively. Relative entropy is also known as Kullback Leibler distance. To utilize this relationship, the distribution of each image is required. These distributions are modeled by the gray level histogram of the image. The resulting $D(p//q)$ value is then compared to a threshold for change determination. The threshold is selected manually and is highly dependent on image dynamic range. Since spatial information is thrown away during the calculation of the histogram, the use of the marginal densities makes the metric insensitive to type II changes. To address this problem we, consider the use of the joint densities because these densities maintain spatial information. This leads to the formulation of a new detection metric relative

Change type	Original object			Change object		
	x, y	size	Intensity	x, y	size	Intensity
III	205, 205	10×10	100	205, 205	10×10	100
III	100, 100	10×10	100	100, 100	10×10	100
II	50, 435	10×10	100	58, 426	10×10	100
I	250, 250	10×10	100	250, 250	10×10	115
none	135, 333	-	-	135, 333	-	-

Table 5.1: Configurations of change blocks in phantom.

	q_c	q_d	\hat{q}_c	\hat{q}_d	GRE	AHST	Chi
1	+300	0	302	0	10.8	.05	.2565
2	0	12.04	9.89	15.86	4.99	.2	.3
3	0	0	0	0	2.87	0	0
4	100	0	-100	N/A	2.56	.0013	.00319
5	N/A	N/A	-	-	0	0	0

Table 5.2: Change Quantification results

entropy.

$$D(p_{xy} // p_{xx}) = \sum p_{xy} \log \frac{p_{xy}}{p_{xx}}$$

This metric measures the inefficiencies of assuming that p_{xx} is the distribution for p_{xy} .

The next phase of processing is change quantification. In this process, the characteristics of the change are determined (i.e. amount, shape, change). This is performed in a multistep process. First, segment the image into two classes. Second, compare the segmentation image with the reference segmented image. Third, form objects from each image and calculate object shape area and center of gravity. Finally, calculate the object overlap and size of difference. The results are then stored in the site model for the next stage of processing.

5.3 Simulation Experiments

To simulate this portion of the system, a phantom mammogram sequence containing four manually changed regions was processed. The three types of change were simulated by modifying a $N \times N$ block of manually changed pixels. Table 5.1 shows the four different configurations. To make the blocks more natural, Gaussian filters are applied to smooth out the edges. To isolate the change detection performance, the phantom sequence was assumed to be perfectly registered. This is accomplished by using the same mammogram in both images of the sequence. We further assume that the radiologist has identified the regions of interest, a 30×30 block of pixels, a pori. Generally, in most change detection metrics a function is evaluated yielding a value which is then compared to a threshold. For this simulation it is assumed the detection threshold is predetermined at 0.5. The performance of joint global relative entropy (GRE) will be compared to two video sequence metrics, an absolute histogram (AHST) and chi square metric (CHI). The quantification portion will be tested by quantitative comparison of the phantom blocks.

Table 5.2 contains the results from processing the phantom where q_c and q_d are the true Δ area and location respectively; and \hat{q}_c and \hat{q}_d are the estimated Δ area and location. For the detection phase of processing we see that GRE metric obtains favorable detection results on all three types of change. The GRE values are $>>$ than the threshold. This indicates that possibly the threshold can be increased which would improve robustness by decreasing the possibility of noise being flagged as change. On the other hand, AHST and CHI fail to detect change at all. This is attributed to the dependence of these metrics on the marginal densities which do not store spatial information. The values produced by these two metrics are $<<$ than the threshold. One would tend to think that performance for these metrics could be improved by decreasing the threshold, but this would only serve to flag noise differences as change. The superiority of the GRE metric can also be seen by examining the ranges of values. The GRE ranges from 0..10.8, while AHST and CHI teams range from 0..0.3 and 0..0.2 respectively. These ranges can also be called dynamic range (value ranges). In communication systems dynamic range is a indicator to the systems sensitive. This same ideal applies to the detection metrics. The GRE metric has a larger spread than AHST and CHI which allows it to capture more and smaller amounts of change.

In the quantification phase, the algorithm accurately quantifies type III change. In this example, the true area difference was 300 pixels². The estimated area difference was 302 pixels². In this case, the translation was estimated

with exactly 0 pixels. In the type II change example the areas remained the same, but a translation of 12 pixels was recorded. The algorithm estimated an area change of 9 pixels² and a translation of 15 pixels. The error in the area could be attributed to the inability of the object selection process to extract the object. Generally, this occurs when the block is the same intensity level as the background. In type I change, the algorithm estimates 0 area change and 0 translation. To fully test the algorithm, an example was selected where no change occurred at all. These results are shown on the bottom row of Table 5.2. Here we see that GRE, AHST, and CHI did not flag this region as changed, but it is difficult to tell if AHST and CHI really found no change or are producing values in their dynamic range.

Chapter 6

Experimental Results and Discussion

6.1 Introduction

The main objective of this research is to detect biological change in a temporal sequence of mammograms. Different types of change can occur between mammograms acquired overtime. The first type of change is natural change which includes weight change and tissue composition change. The next type of change is image acquisition change. This includes the changes caused by breast positioning, breast compression, and differences in imaging equipment. Finally, change that possibly indicates cancer or the onset of cancer. This type is usually visualized as a microcalcification or mass [3]. The first two types of change generally affect the complete image and are classified as global change. On the other hand, the third type of change is usually localized to a region and is classified as local change. Due to the enormous number of combinations relating to the first two types of change, we focus attention on local change. In addition, we also only consider change calculated from a radiologist selected localized window. Local change has been shown to be an indicator of the onset of cancer [4]. Currently, radiologists perform change analysis manually following a specific procedure [3]. Automation of this task could help to reduce the fatigue felt by the radiologists which may lead to an increase in analysis accuracies. This chapter presents and discusses the results generated by applying the developed change detection algorithm to real mammogram sequences. See Figure 6.1 for a system overview and flow diagram. Next, the results of several example mammogram sequences will be discussed.

6.2 Experiment Results and Discussion

The first example is a sequence composed of two right CC views of the same patient acquired on 1/21/93 and 2/3/99 as shown in Figure 6.2 a and b. The image acquired on 2/3/99 contains a suspicious region located at (77, 317). Figure 6.2a is taken as the reference image and used to construct the site model. The users input to the site model is the region of interest, which is a 30×30 square centered around the point (77, 317). The radiologist selects the window size manually as seen in Figure 6.3. After construction of the site model, processing new images can commence. The first step is the extraction of parameters used in initial registration. This includes objects and their descriptions. Next, multi-object PAR is performed using 2 of the objects as seen in Figure 6.4. The resulting initial registration pair is shown in Figure 6.5. Comparing Figure 6.5 and Figure 6.2 we see that most of the scale difference between the images has been corrected. Finer alignment could be obtained if control points were known. Using the initially registered image, final registration parameters are extracted. These parameters include potential control points and their associated signatures. Next, the recently extracted potential control points are matched with the potential control points from the site model to obtain the final control points. This matching is performed by two methods in this research. Figure 6.6 shows control points obtained by matching signatures using the Pearson correlation coefficient while Figure 6.7 shows control points obtained by matching Nearest Neighbor. In this example, Pearson matching yields 13 control point pairs out of a pool of 66 potential control points or a match rate of 0.197. This rate is low because the deformation between the site and incoming image produced different potential control point pools in each image. Thus, signature matching yields few matches when signature correlation is low. The final control points in this example are clustered into 2 loose groups located on the top and bottom of the breast. This appears to be caused by the existence of dense tissue near the center of the breast. In dense tissue, the monotony operators (used to find elongated structures) appear to have problems when the tissue intensities are nearly constant. Nearest Neighbor matching, on the other hand, yielded 27 control points evenly distributed across the image. This yields a match rate of 0.409. This number is still low, but more acceptable. Both matching rates could be improved by the increase in the localized search window size, but the probability of mis-match would also increase. Mis-match control points cause gross distortion in the transformed image. Since our method of control

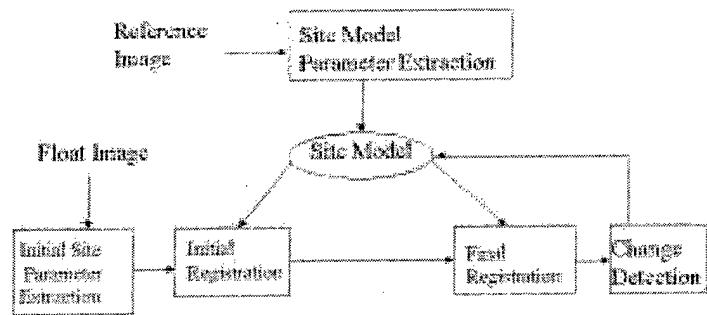


Figure 6.1: Change detection process flow

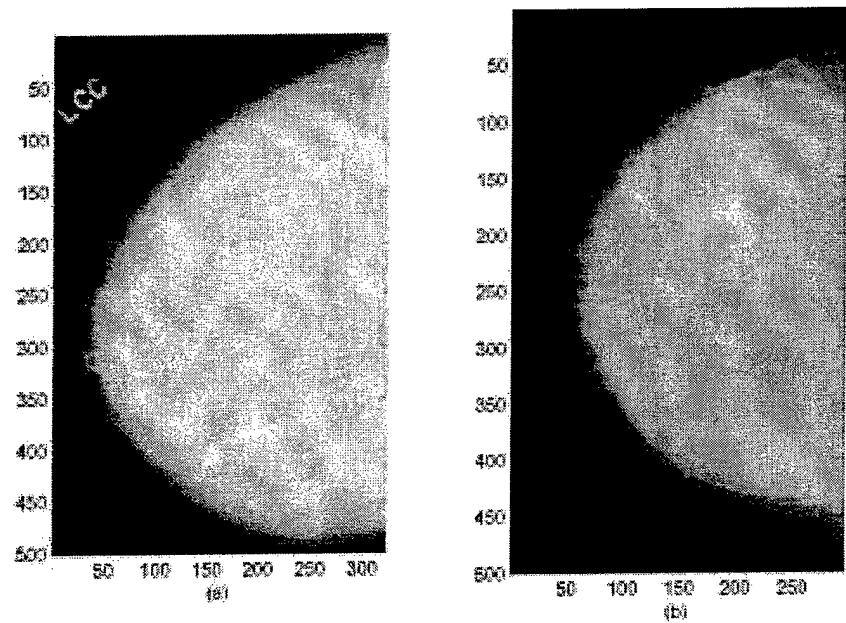


Figure 6.2: Raw mammogram sequence. (a) 1/21/93. (b) 2/3/99.

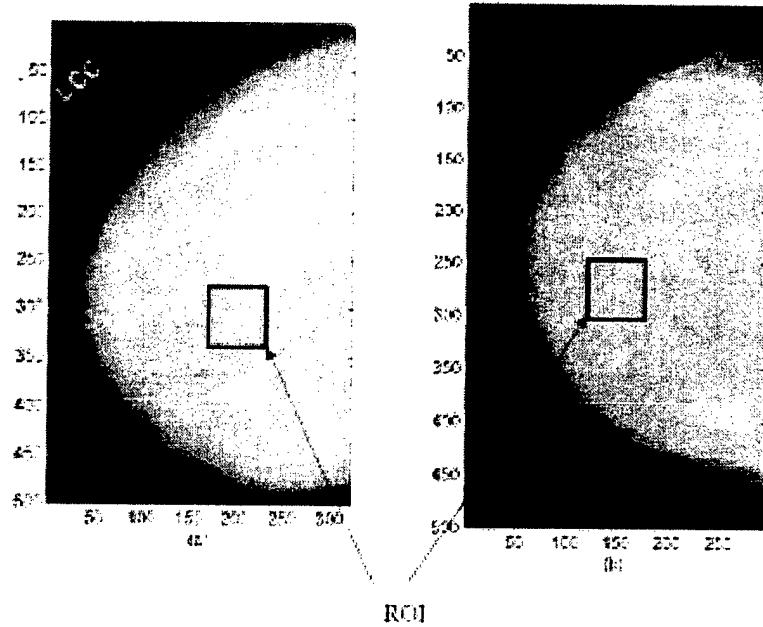


Figure 6.3: Marked region of interest.

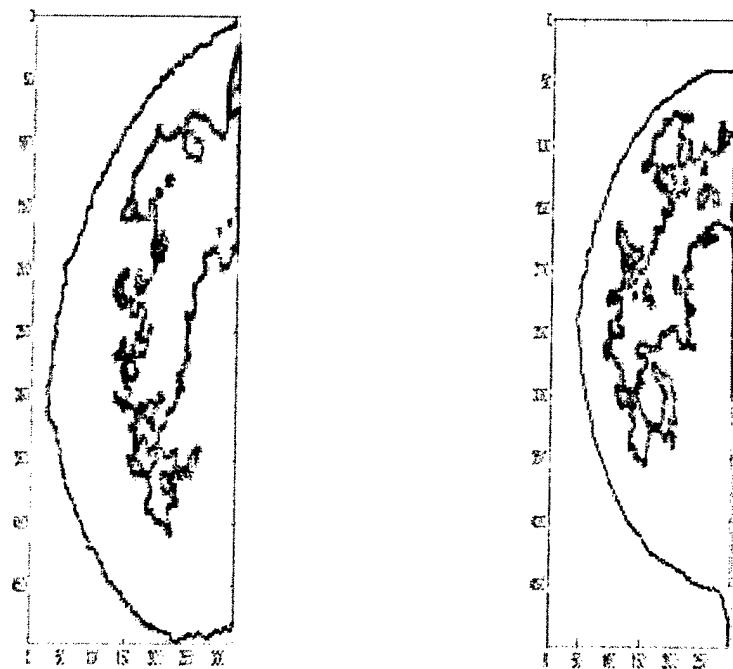


Figure 6.4: Objects used in Multi-object transform.

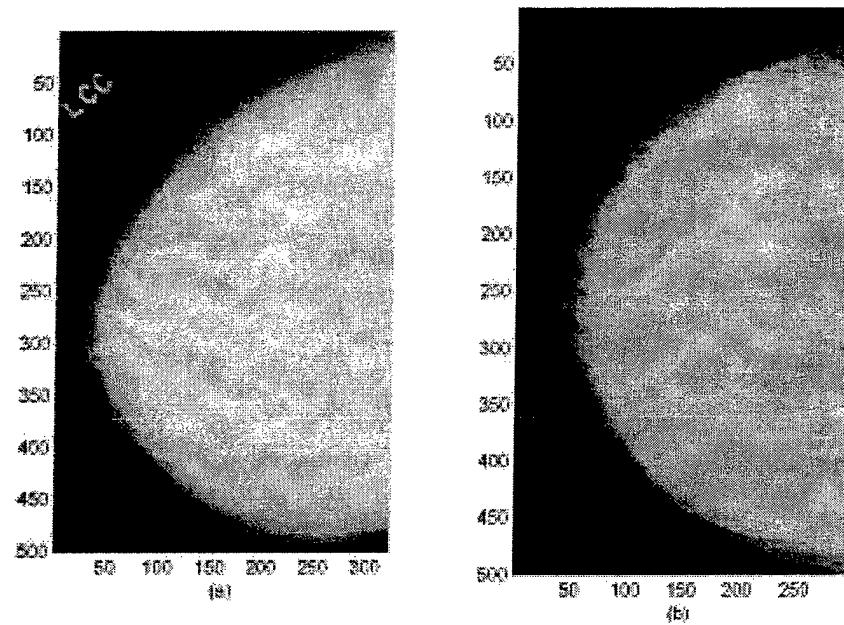


Figure 6.5: Multi-object PAR image pair.

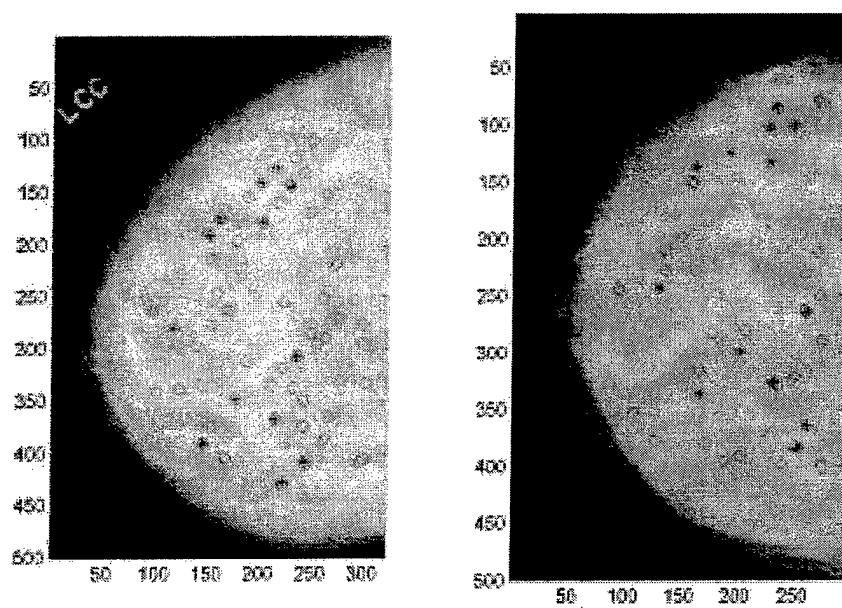


Figure 6.6: Potential 'o' and final '*' control points using Pearson correlation.

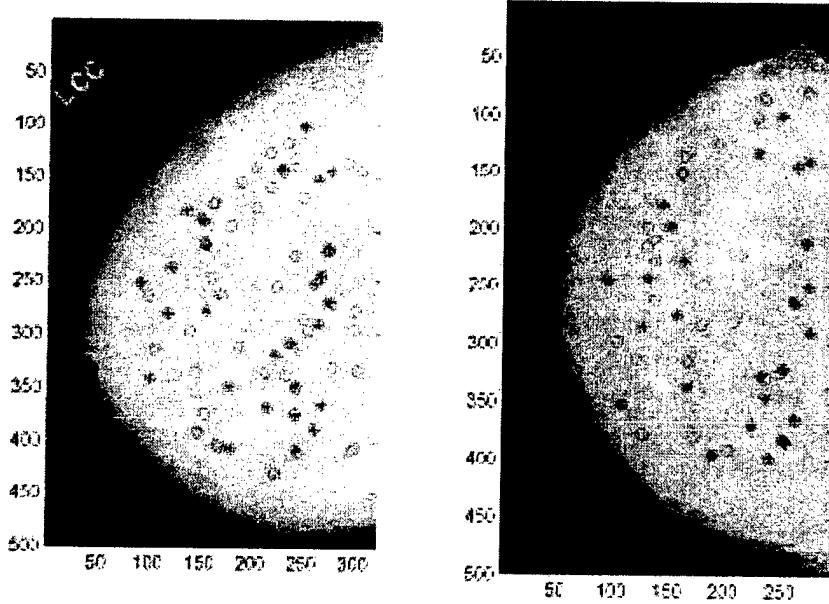


Figure 6.7: Potential */o/* and final */*/* control points using Nearest Neighbor method.

point extraction is based on [58] we suffer the same image dependence problems as [58]. Window sizes, thresholds, and monotony operator dimensions are among the key parameters that need to be adjusted on a per image basis. For our research we use a window size of 10×10 , threshold of 6, monotony operator dimension of 1 and 5. These values were experimentally determined using visual inspection of initial output. Next, the final transform is derived and applied to the image pixel by pixel resulting in the pair shown in Figure 6.8.

To perform change detection, the corresponding region of interest from the incoming image is compared to the site model. The histograms of the two regions are compared in Figure 6.9. From this figure, the difference is visually apparent as the two regions have different distributions. Three change metrics were applied yielding the following results: global relative entropy (GRE) 23.63; absolute histogram difference (AHST) 0.885; and chi square (CHI) 1.0. The last two metrics are video sequence metrics and serve as comparisons of existing change methods. Given the threshold of 1.5 which was determined experimental, both AHST and CHI miss the change which means they appear to be insensitive to slight scene changes, but GRE detects the change. In fact, this change resulted in a GRE value $\gg 1.5$. It would appear that the threshold could be increased, but this would increase the probability of miss.

Unlike the phantom studies performed in the other chapters, no ground truth exists for quantification of the changed region. For this reason, visual inspection is used to examine the results. The quantification process determined an area difference of 353 pixels which was verified by an radiologist during a manual inspection. The detected area is larger then the area estimated by the radiologist because the object extraction process cannot remove all of the background pixels. 54 out of the 354 pixels are background pixels.

In the next example, the radiologist identified a suspected area (region of interest) on the final mammogram (i.e. first image). The raw sequence is given in Figure 6.10 and is composed of a right CC view of a patient acquired on 3/5/96 and 2/24/99. The */X/* marks on the image are the location of the change region. On the site image the */X/* is the associated point. For this example, two objects were selected for use with the multi-object PAR. Figure 6.11 is the resulting transformed image where */X/* marks the change location. — control points were matched out of — potential control points to form the TPS transform. The final warped image is shown in Figure 6.12. From examination of the image it appears distortion occurred, but the location of the */X/* on both images appear to visually cover the same portion of tissue. In comparing, Figure 6.10, 6.11, and 6.12 we indeed notice this fact. The image's distorted look is caused by too few control points on the skin line (or region). Thus, the affect of the

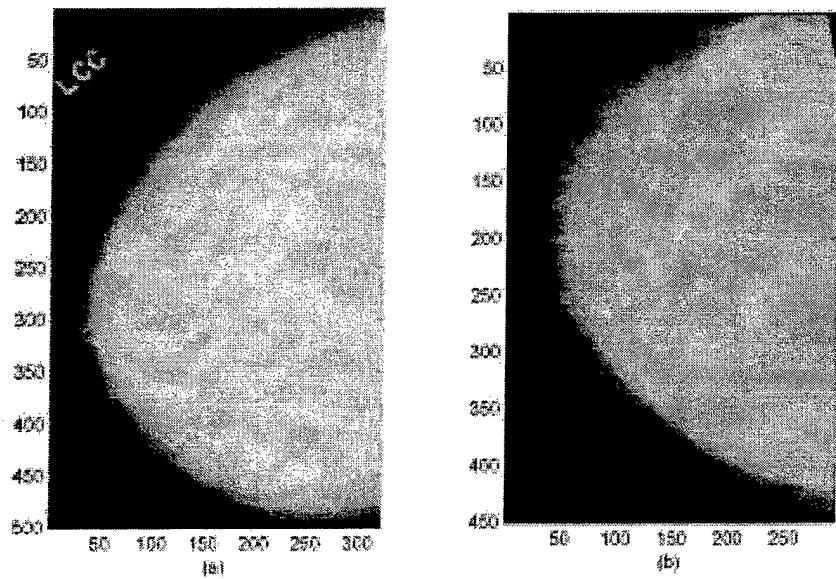


Figure 6.8: Final warped image pair.

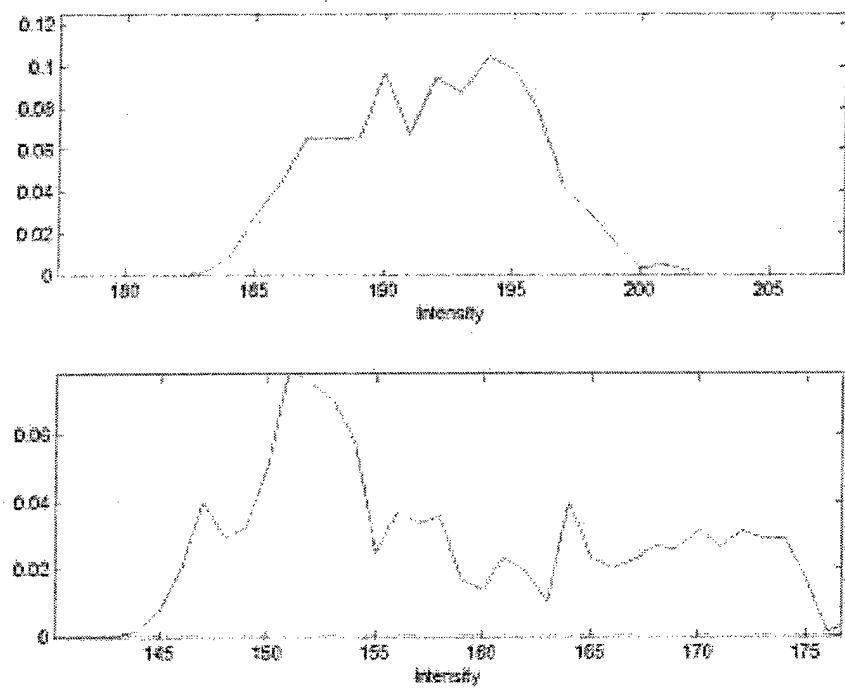


Figure 6.9: Histogram of corresponding regions of interest

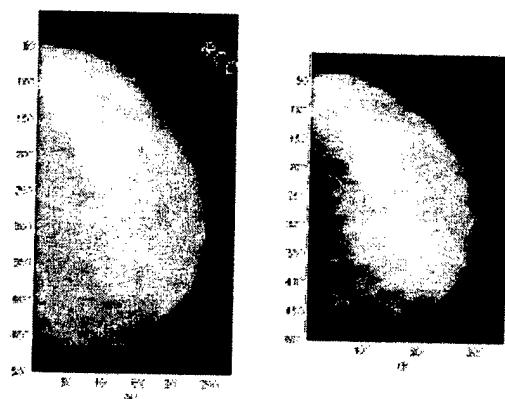


Figure 6.10: (a) Reference image 3/5/96 , (b) float image 2/24/99.

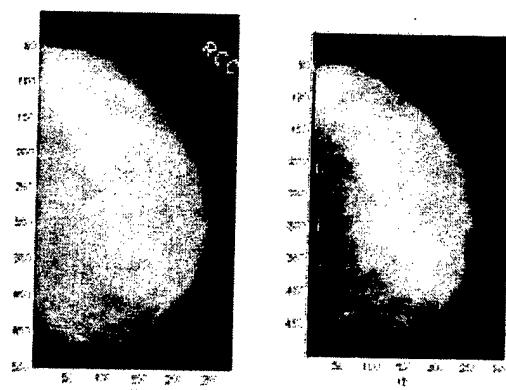


Figure 6.11: (a) Reference image, (b) Multi-object PAR image.

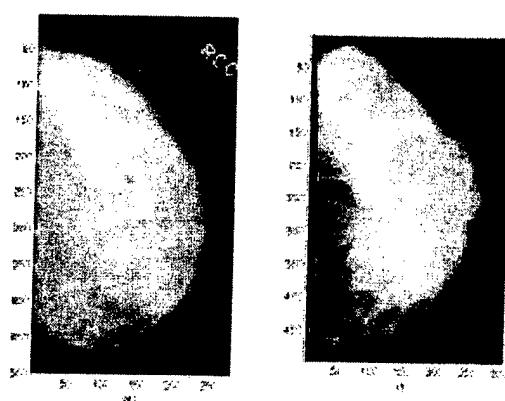


Figure 6.12: Final image pair after registratioin. (a) Reference image. (b) Warped image.

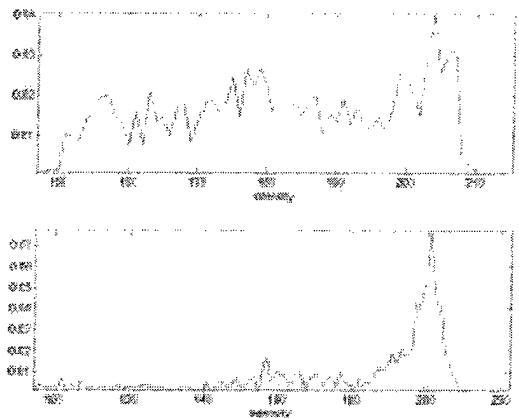


Figure 6.13: Histogram comparison between the two local change windows

control points on the skin-line pixels is greatly reduced causing a massive warping effect. The algorithm was then used to see if the area was present on the site mammogram. The intensity histograms of two regions are shown in Figure 6.13. From here change can be visually determined. To detect the change, GRE, AHST, and CHI metrics were calculated yielding the following values 22.9, 0.512, and 0.4611 respectively. Again, the GRE metric is \gg than the threshold while AHST and CHI fall below the metric. The quantification results estimate a 530 pixels² change. The true change is closer to 9 pixels². The massive error results from the inability to extract the object from the background of similar pixel intensities resulting in a large selected region.

6.2.1 Summary

Change detection not only highlights existence of possible changed regions, but when combined with the site model provides a patient history by showing site progression. One of the key components of change detection is image registration. In this chapter, we applied our multi-step registration algorithm to mammogram sequences. Acceptable registration and change detection were obtained. Improvement in control object selection and control point extraction would go along way to improving the overall results. The key to registration is landmarks between the images. In this research, we use objects and points as landmarks. Current methods of object and point selection are image dependent and adhoc. Incorrect assignment of control points/objects could cause erroneous transformation. This change detection is not exact, but would be sufficient to flag a radiologist to review the area. The main results of this study consisted of the automatic alignment of mammograms, detection of change in a local window, and implementation of a mechanism to store and build up patient information via the site model.

Chapter 7

Future Work

In this chapter the future directions of this research are discussed. The clinical problem of change detection in a temporal sequence of mammograms has lead to several interesting and challenging technical problems (non-rigid registration and change detection). Of these problems, performing image registration is the most important. Currently only four approaches, including the approach presented in this research, exist for the registration of mammograms. So, development toward this problem could yield major benefits in the clinical field.

With that in mind, the future direction of this research is two fold, registration and change detection. Several key aspects of registration have bee researched here. They are control point determination and use of multiple objects in registration. Control point determination covers potential control point detection and control point correspondence theories. While multi-object registration, covers object definition, object correspondence, and transform combination. Extension of these ideas could lead to more robust registration which ultimately leads to better change detection.

In terms of change detection, the future work includes extension of change detection quantification theories to more accurately categorize the change. This includes more automatic object detection and shape description. Another change detection direction is the expansion of this local change approach to consider global change detection and quantification. This will include discrimination between natural and unnatural change and change localization. Possible signal and image processing techniques include wavelets and statistical based processing.

**Enhance nipple detection using approach in NIPPLE PAPER

Use CC and MLO to create site model *issue obtain correspondence between images and points
control point correspondence (HMM, FD, etc.)****

Chapter 8

Appendix A: Information Criterion

Determining the number of components in a mixture signal is useful in numerous applications from speech processing to object recognition. These type of problems are termed model selection or cluster validation in the literature [23]. The main goal in these type of problems is to estimate, given the data, the number of components K , are present in the mixture signal. This is accomplished by evaluating a function (Information Criterion IC) for reasonable values of K . \hat{K} is taken as the K value that yields the minimum function result. The first and most widely used IC is Akaike Information Criterion (AIC).

8.1 Theory

The AIC formulation can be derived using the following model [23]. Suppose our data is represented by N random vectors given by $Y = \{y_1, \dots, y_N\}$. Further assume that the distribution of y is composed of K components where the distribution of the k^{th} component is $f_k(Y/\theta_{ml}^k)$ where θ_{ml} are the ML estimate of the features. So the goal of the IC is to find the K that maximize the function. Since we assume our distribution is a Gaussian, finding its maximum is equivalent to minimizing the log of the distribution function. The results are the AIC equations given below.

$$AIC(K) = -2 \log(f(x/\phi_{ml})) + 2 * K_a \quad (8.1)$$

$$K' = \arg \min AIC(K); 2 \leq K \leq K_0 \quad (8.2)$$

where $f(x/\phi_{ml})$ is the conditional likelihood function distribution given the maximum likelihood feature vector ϕ_{ml} . K_a is the number of free parameters to estimate and was added to make the AIC estimate an unbiased estimate of the mean distance between $f(x/\theta)$ and $f(x/\theta')$ where θ' is the estimated parameter vector.

8.2 Simulation Experiments

To illustrate this algorithm two examples were processed a four class phantom shown in Figure 8.1 and a real mammogram. For each example, the k ranged from 2..10. Figure 8.2 shows the plot the AIC curve for the phantom and Figure 8.3 shows the plot for the mammogram. From these plots we see that \hat{K} is 4 and \hat{K} is 8. The results correspond to results achieved in [27].

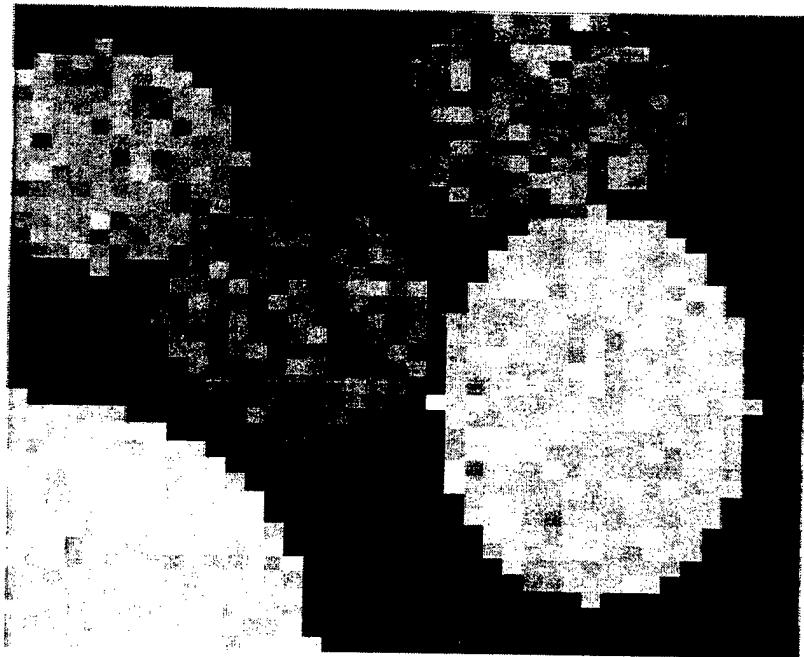


Figure 8.1: Four class phantom

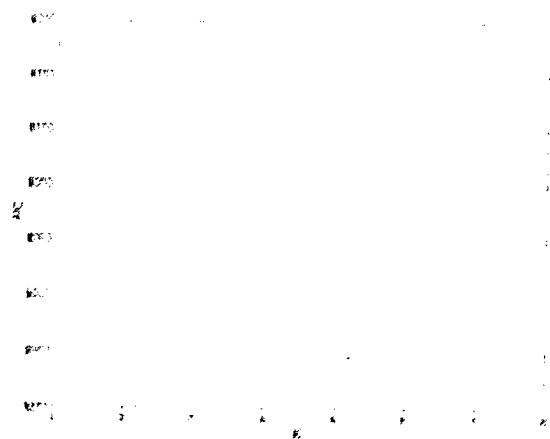


Figure 8.2: AIC plot of four class phantom

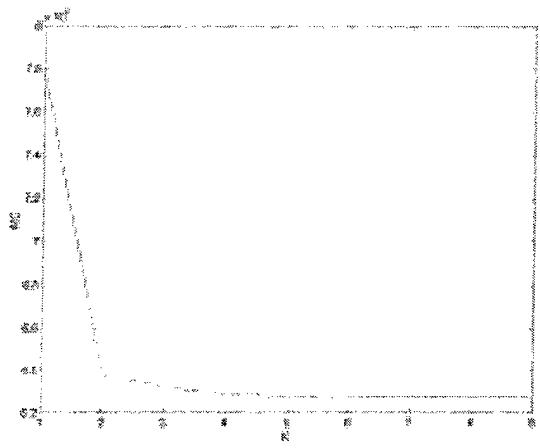


Figure 8.3: AIC plot of mammogram image.

Bibliography

- [1] D. B. Kopans, *Breast Imaging*, J.B. Lippincott Co., Philadelphia, Pa., 1989.
- [2] D. Dance, "Physical Principles of Breast Imaging," *Proceedings of the 3rd International Workshop on Digital Mammography*, pp. 11-18, Chicago, IL, 1996.
- [3] P B. Dean, "Overview of Breast Cancer Screening," *Proceedings of the 3rd International Workshop on Digital Mammography*, pp. 19-31, Chicago, IL, 1996.
- [4] G. Cardenosa, "Mammography: An Overview," *Proceedings of the 3rd International Workshop on Digital Mammography*, pp.3-10, Chicago, IL, 1996
- [5] M. Sallam and K.W. Bowyer, "Registering Time Sequences of Mammograms Using Two-dimensional Image Unwarping Technique," *Proceedings of the 2nd International Workshop on Digital Mammography*, pp. 121-130, York England, 1994.
- [6] N. Vujovic, P. Bakic, and D. Brzakovic, "Detection of Potentially Cancerous Signs by Mammogram Follow-up," *Proceedings of the 3rd International Workshop on Digital Mammography*, pp. 421-424, Chicago, IL, 1996
- [7] N. Vujovic and Brzakovic, "Establishing the Correspondence Between Control Points in Pairs of Mammographic Images," *IEEE Trans. on Image Processing*, vol. **6**, pp. 1388-1399, 1997.
- [8] W. K. Zouras, et. al., "Investigation of a Temporal Subtraction Scheme for Computerized Detection of Breast Masses in Mammograms," *Proceedings of the 3rd International Workshop on Digital Mammography*, pp. 411-415, Chicago, IL, 1996.
- [9] M. A. Wirth and C. Choi, "Multimodal Registration of Anatomical Medial Images," *Australian Pattern Recognition Society, Conference on Image processing*, Oct. 1996.
- [10] R.M. Ford, et. al., "Metrics for Scene Change Detection in Digital Video Sequences," *Multimedia Computing and Sys. 97 Proceedings IEEE International Conference on*, pp. 610-611, 1997.
- [11] X. Dai, S. Khorram, "The Effects of Image Misregistration on the Accuracy of Remotely Sensed Change Detection," *IEEE Trans. Geosci. Remote Sensing*, vol. **36**, pp.1566-1577,1998.
- [12] P. Dhawan, et. al., "Iterative Principal Axes Registration Method for Analysis of MR-PET Brain Images," *IEEE Trans. Biomed. Eng.*, vol. **42**, pp. 1079-1087,1995.
- [13] C. L. Lin, Q. Zheng, R. Chellappa, L. S. Davis, X. Zhang, "Site model supported monitoring of aerial images," *Computer Vision and Pattern Recognition*, June pp. 694-700,1994.
- [14] L. Ott, *An Introduction to Statistical Methods and Data Analysis*, pp. 220-229, Wadsworth Pub. Co.,Belmont, Ca., 1977.
- [15] R. C. Jain, et. al., "On the Analysis of Accumulative Difference Pictures from Image Sequences of Real World Scenes," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. **1**, pp. 206-213, 1979.
- [16] R.L. Campbell, "Image Enhancement via morphological Filtering," *International Conference on Signal Processing Applications & Technology*, vol. **2**, pp. 1133-1137, Boston Ma. 1996.
- [17] Z. Liang, "Tissue Classification and Segmentation of MR Images," *IEEE Eng. in Medicine and Biology*, pp. 81-85. Mar. 1993.
- [18] Y. Wang, "MRI statistics and model-based MR image analysis," Ph.D. report, University of Maryland Graduate School, Baltimore, MD, April 1995.

- [19] T. Lei, W. Sewchand, "Statistical Approach to X-Ray CT Imaging and Its Applications in Image Analysis — Part II: A New Stochastic Model-Based Image Segmentation Technique for X-Ray CT Image," IEEE Trans. Med. Imaging, vol.11, no. 1, pp. 62-69, Mar. 1992.
- [20] Y. Wang, T. Adali, S.B. Lo, "Automatic Threshold Selection Using Histogram Quantization," J. Biomedical Optics, Vol. 2, No. 2, pp. 211-217, April 1997.
- [21] C. A. Bouman, M. Shapiro, "A Multiscale Random Field Model for Bayesian Image Segmentation," IEEE Trans. Image Proc., Vol.3, No. 2, pp. 162-176, Mar. 1994.
- [22] J. Zhang, J.W. Modestino, D. A. Langan, "Maximum-Likelihood Parameter Estimation for Unsupervised Stochastic Model-Based Image Segmentation," IEEE Trans. Image Proc., Vol.3 No. 4, pp. 404-420, July 1994.
- [23] D.A. Langan, J.W. Modestino, J. Zhang, "Cluster Validation for Unsupervised Stochastic Model-Based Image Segmentation," IEEE Trans. Image Proc., Vol. 7 No. 2, pp. 180-195, Feb. 1998.
- [24] Z. Liang, J.R. MacFall, D. P. Harrington, "Parameter Estimation and Tissue Segmentation from Multispectral MR Images," IEEE Trans. Medical Imaging, Vol. 13, No. 3, pp. 441-449, Sept. 1994.
- [25] Dempster, A. P., Laird, N.M. and Rubin, D. B., "Maximum Likelihood from Incomplete Data via the EM algorithm," J. Roy. Soc. Statist., B., No. 1, pp. 1-38, 1977.
- [26] Y. Wang, T. Adali, M.T. Freedman, and S. K. Mun, "MR Brain Image Analysis by Distribution Learning and Realization Labeling," Proc. 15th South. Biomed. Eng. Conf., pp. 133-136, Dayton Ohio, Mar. 1996.
- [27] Y. Wang, T. Adali, S-Y Kung, and Z. Szabo, "... A Probabilistic Neural Network Approach," IEEE Trans. Image Proc., Vol. 7, No. 8, pp.1165-1181, Aug. 1998.
- [28] Y. Wang, T. Lei, " A New Look at Finite Mixture Models in Medical Image Analysis," ISSIPNN, 1994, pp. 33-35.
- [29] S. M. LaValle, S. A. Hutchinson, "A Bayesian Segmentation Methodology for Parametric Image Models," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 17, No. 2, Feb. 1995, pp 211-217.
- [30] M. Cheriet, J.N. Said, C. Y. Suen,"A Recursive Thresholding Technique for Image Segmentation," IEEE Trans. Image Processing, Vol. 7 No. 6, June 1998, pp. 918-921.
- [31] S. L. Sclove, " Application of the Conditional Population-Mixture Model to Image Segmentation," IEEE Trans. PAMI, Vol. PAMI-5, No. 4, July 1983, pp. 428-433.
- [32] J. Zhang, J. W. Modestino, "A Model-Fitting Approach to Cluster Validation with Application to Stochastic Model-Based Image Segmentation," IEEE Trans. PAMI, Vol. 12, No. 10, Oct. 1990, pp. 1009-1017.
- [33] Y. Delignon, A. Marzouki, W. Pieczynski,"Estimation of Generalized Mixtures and Its Application in Image Segmentation," IEEE Trans. Image Processing Vol. 6, No. 10, Oct. 1997, pp. 1364-1375.
- [34] S. S. Saquib, C. A. Bouman, K. Sauer,"ML Parameter Estimation for Markov Random Fields with Application to Bayesian Tomography," IEEE Trans. Image Processing, Vol. 7, No. 7, July 1998, pp.1029-1044.
- [35] C. Bouman, B. Liu,"Multiple Resolution Segmentation of Textured Images," IEEE Trans. PAMI, Vol. 13, No.2, Feb. 1991, pp. 99-113.
- [36] K. Held et al., "Markov Random Field Segmentation of Brain MR Images," IEEE Trans. Medical Imaging, Vol.16, No. 6, Dec. 1997, pp. 878-886.
- [37] T.M Chang, Y.H Liu, C.H. Chen, et al."Intermodality Registration and Fusion of Liver Images for Medical Diagnosis," Intelligent Information Systems 1997 ISS 1997 Proceedings.
- [38] T.D. Zuk, M.S. Atkins, "A Comparison of Manual and Automatic Methods for Registering Scans of the Head," IEEE Trans. Medical Imaging, Vol. 15, No. 5, Oct. 1996 pp 732-744.
- [39] WM. Wells III, P. Viola, H. Atsumi, et al., "Multi-modal volume registration by maximization of mutual information," Medical Image Analysis Vol. 1, No. 1, pp35-51.
- [40] A. Moskalik, P.L. carson, C.R. Meyer, J.B. Fowlkes, J.M. Rubin, et al., "Registration of Three-Dimensional Compound Ultrasound Scans of the Breast for Refraction and Motion correction," Ultrasound Med. & Biol., Vol. 21 No. 6. pp 769-778, 1995.

- [41] P. A. Van den Elsen, J.B. Antoine Maintz, et al., "Automatic Registration of CT and MR Brain Images Using correlation of Geometrical Features," IEEE Trans. Medical Imaging Vol. 14 No. 2, June 1995, pp. 384-396.
- [42] C. R. Maurer Jr., G. B. Aboutanos, et al."Registration of 3-D Images Using Weighted Geometrical Features," IEEE Trans. Medical Imaging, Vol. 15, Dec. 1996, pp. 836-849.
- [43] P. A. Vand den Elsen, E. D. Pol, Max A. Viergever," Medical Image Matching – A review with Classification," IEEE Eng. Medicine and Biology, Mar. 1993, pp. 26-38.
- [44] C.R. Maurer Jr., J.M. Fitzpatrick, et al."Registration of Head Volume Images Using Implantable Fiducial Markers," IEEE. Tra s. Medical Imaging Vol. 16 No. 4, Aug. 1997, pp. 447-462.
- [45] C. Davatzikos, J. L. Prince and R.N. Bryan, " Image Registration Based on Boundary Mapping," IEEE Trans. Medical Imaging, Vol. 15, No. 1 Feb. 1996, pp. 112-115.
- [46] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality Image Registration by Maximization of Mutual Information," IEEE Trans. Medical Imaging Vol. 16, No. 2, April 1997, pp. 187-198.
- [47] R. J. Althof, M. G. J. Wind, J. T. Dobbins," A Rapid and Automatic Image Registration Algorithm with Subpixel Accuracy," IEEE Trans. Medical Imaging Vol. 16. No. 3, June 1997, pp. 308-316.
- [48] C. R. Maurer, Jr., et al., " Registration of Head CT Images to Physical Space Using a Weighted Combination of Points and Surfaces," IEEE Trans. Medical Imaging , Vol. 17. No 5. Oct. 1998, pp. 753-761.
- [49] C.A. Pelizzari et al., " Comparison of Two Methods for 3D Registration of PET and MRI Images," AIC of IEEE Eng. in Medicine and Biology Society, Vol. 13, No. 1, 1991 pp. 221-223.
- [50] L. K. Arata, A. P. Dhawan," Iterative Principal Axes Registration: A New Algorithm for Retrospective Correlation of MR-PET Brain Images," AIC of IEEE Eng. in Medicine and Biology Society, Vol. 14, No. 7, 1992 pp. 2776-2778.
- [51] S. C. Strother et al., " Quantitative Comparisons of Image Registration Techniques Based on High-Resolution MRI of the Brain," Journal of Computer Assisted Tomography, Vol. 18, No. 6, Nov/Dec. 1994, pp. 954-962.
- [52] M. S. Brown et al., " Method for Segmenting Chest CT Image Data Using an Anatomical Model: Preliminary Results," IEEE Trans. on Medical Imaging Vol. 16 No. 6 Dec. 1997, pp. 828-839.
- [53] L. K. Arata et al., "Three-Dimensional Anatomical Model-Based Segmentation of MR Brain Images Through principal Axes Registration," IEEE Trans. on Biomedical Engineering, Vol. 42. No. 11, Nov. 1995, pp. 1069-1078.
- [54] K. Woods, J. Wang, M. T. Freedman, " Unsupervised Tissue Quantification and Segmentation from 3-D MRI Brain Images", IASTED SIP 1998. pp. 772-775.
- [55] D.L. Collins, A.P. Zijdenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes, A.C. Evans : "Design and Construction of a Realistic Digital Brain Phantom" IEEE Trans. on Medical Imaging, Vol.17, No.3, p.463-468, June 1998.
- [56] R.K.-S. Kwan, A.C. Evans, G.B. Pike : "An Extensible MRI Simulator for Post-Processing Evaluation" Visualization in Biomedical Computing (VBC'96). Lecture Notes in Computer Science, Vol. 1131. Springer-Verlag,1996. 135-140.
- [57] C.A. Cocosco, V. Kollokian, R.K.-S. Kwan, A.C. Evans : "BrainWeb: Online Interface to a 3D MRI Simulated Brain Database" NeuroImage, Vol.5, No.4, part 2/4, S425, 1997 – Proceedings of 3rd International Conference on Functional Mapping of the Human Brain Copenhagen, May 1997.
- [58] N. Vujovic, "Registration of Time-Sequences of Random Textures with Application to Mammogram Follow-up," Ph.D. report, Lehigh University, May1997.
- [59] N. S. Vujovic, et. al., "Analogic Algorithm for Point Pattern Matching with Application to Mammogram Followup," 4th Workshop on Cellular Neural Networks and App., June 24-28, 1998.
- [60] D. Brzakovic, et. al., " Mammogram Analysis by Comparison with Previous Screenings," Proc. of the 2nd International Workshop on Digital Mammography, York, England, pp. 131-139,1994.
- [61] N. Vujovic and D. Brzakovic," Feature Point Identification and Regional Registration in Sequences of Non- Structured Texture Images," Proc. International Conference on Image Proc., vol. 3, pp. 156-159, 1995.

- [62] D. Brzakovic, et. al., "Early Detection of Cancerous Changes by Mammogram Comparison," Proc. SPIE Visual Communications and Image Processing 1994, pp. 1520-1531, Chicago, IL, 1994.
- [63] M. Sallam and K. Bowyer, "Detecting Abnormal Densities in Mammograms by Comparison to Previous Screenings," Proc. of the 3rd International Workshop on Digital Mammography, Chicago, IL, pp. 417-420, 1996.
- [64] M. Sallam, et. al., "Screening Mammogram Images for Abnormalities Developing Over Time," Proc. IEEE Nuclear Science Symposium and med. Imaging Conference, pp. 1270-1272, 1992.
- [65] M. Abdel-Mottaleb, et. al., "Locating the Boundary Between the Breast Skin Edge and Background in Digitized Mammograms," Proc. of the 3rd International Workshop on Digital Mammography, Chicago, IL, pp. 467-470, 1996.
- [66] R. Chandrasekhar and Y. Attikiouzel, "A Simple Method for Automatically Locating the Nipple on Mammograms," IEEE Trans. Med. Imaging, vol. 16, no. 5, pp. 483-494, 1997.
- [67] F. L. Bookstein, "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations," PAMI vol. 11, no. 6, pp. 567-585, June 1989.
- [68] M. A. Wirth, et. al., "Point to Point Registration of Non-Rigid Medical Images Using Local Elastic Transformation Methods," IEE. Image proc. App. July 14-17 1997.
- [69] M. A. Wirth, et. al., "A Nonrigid-Body Approach to Matching Mammograms," 7th International conference on Image proc. App., pp 484-488, 1999.
- [70] Goshtasby, "Registration of Images with Geometric Distortions," IEEE Trans. Geo. Remote Sensing, vol. 26, no. 1, pp. 60-64, Jan. 1988.
- [71] G.J. Ettinger, W.E.L. Gunson, et al. "Automatic registration for Multiple Sclerosis Change Detection" Proceeding of the IEEE Workshop on biomedical Image Analysis, Seattle, WA. 1994.
- [72] B. M. Hemminger, et. al., "Evaluation of Digital Processing Methods for the Display of Digital Mammography," SPIE Conf. Image Display, SPIE vol. 3658, pp. 382-393, Feb. 1999.
- [73] L. Brown, "A survey of image registration techniques", ACM Computing-Surveys, vol. 24. pp325-376, New York, 1992.
- [74] R. Collins, et al., "Model matching and extension for automated 3D site modeling," Proc. ARPA Image Understanding Workshop, Washington D.C., pp. 197-204, April 1993.
- [75] R. Collins, A. Hanson, E. Riseman, "Site model acquisition under the UMass RADIUS Project," Proc. ARPA Image Understanding Workshop, Monterey, CA., pp. 351-358 Nov. 1994.
- [76] H.L. Van Trees, "Detection, Estimation, and Modulation Theory," John Wiley and Sons, New York, 1968
- [77] S. J. Orfanidis, "Optimum Signal Processing an Introduction," Mc Graw-Hill Pub. Company, New York, 1988.
- [78] Q. Zheng, R. Chellappa, "A computational vision approach to image registration." IEEE Trans on Image Processing, vol. 2, pp. 311-326, 1993.
- [79] R. Chellappa, et. al., "Site-Model-Based Monitoring of Aerial Images," Proc. ARPA Image Understanding Workshop, pp. 295-318, 1994.
- [80] X. Zhang, et. al., "Automatic Image to Site Model Registration," Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing, pp. 2164-2167, Atlanta, GA, May 1996.
- [81] H. Yildirim, et. al., "Temporal Change Detection by Principal Component Transformation," IEEE International Geoscience and Remote Sensing Symposium, vol. 2, pp. 1227-1229, 1995.
- [82] L. Bruzzone, S. B. Serpico, "An Iterative Technique for the Detection of Land-Cover Transitions in Multi-temporal Remote-Sensing Images," IEEE Tans. on Geoscience and Remote sensing, Vol. 35, no.4, pp. 858-866. 1997.
- [83] T. Yamamoto, et. al., "A Change Detection Method for Remotely Sensed Multi-spectral and Multi-Temporal Images using 3-D Segmentation," IEEE International Geoscience and Remote Sensing Symposium vol. 1, pp 77-79, 1999.

- [84] J.B. Antoine Maintz ,M. A. Viergever, "A survey of medical image registration," Medical Image Analysis, vol.2, pp. 1 – 36, 1998.
- [85] A. K. Jain, "Fundamentals of Digital Image Processing," Prentice-Hall, 1989.
- [86] B. Kolman, "Introductory linear algebra with applications," Macmillan Publishing Co., 1988.
- [87] K. I. Laws, "Rapid Texture Identification," Proc. SPICE Conf. Image Processing for Missile Guidance, pp. 376-380, 1980.
- [88] W. F. Good, et. al., "Image modification for display of temporal sequences of mammograms," Medical Imaging 2000: Image Display and Visualization, Proceedings of SPIE, Vol. 3976, pp. 174-184, 2000.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel in the order listed for Form Page 2.
Follow the sample format on preceding page for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Wang, Yue (Joseph)	POSITION TITLE Associate Professor of Electrical Engineering (tenured) and Radiology [†]		
EDUCATION/TRAINING (<i>Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.</i>)			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
Jiao Tong University	B.S.	1984	Electrical Engineering
Jiao Tong University	M.S.	1987	Electrical Engineering
University of Maryland	Ph.D.	1995	Electrical Engineering
Georgetown University	Postdoc	1996	Radiology

A. Positions and Honors**Positions and Employment**

- 1996-2001 Assistant Professor of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC
 2000- Adjunct Associate Professor of Radiology and Radiological Science[†], Johns Hopkins University School of Medicine, Baltimore, MD
 2001- Associate Professor of Electrical Engineering and Computer Science (tenured), The Catholic University of America, Washington, DC

Other Experience and Professional Memberships

- 1994-present Member, IEEE Signal Processing Society
 1996-present Director, Computational Imaging and Informatics Laboratory, The Catholic University of America
 1998 Member, Tau Beta Pi
 1999-present Member, Technical Committee on Neural Networks for Signal Processing (TC-NNSP), IEEE Signal Processing Society
 1999 Peer Review Committee, "Congressional Directed Medical Research Program-CDMRP", Department of Defense
 2000 Program Committee, IEEE Workshop on Neural Networks for Signal Processing, Australia
 2000 Technical Committee, IEEE Workshop on Multimedia Signal Processing, Australia
 2000-present *ad hoc* Study Section, "Innovative Technologies for the Molecular Analysis of Cancer-IMAT" (ZCA1-M1), National Institutes of Health
 2001 *ad hoc* Study Section, "In-Vivo Cellular and Molecular Imaging Centers-ICMICs" (ZCA1-P20/P50), National Institutes of Health
 2001-present Associate Editor, *IEEE Transactions on Information Technology in Biomedicine*
 2002 Guest Editor, Special Issue on Biometric Signal Processing, *European Journal of Applied Signal Processing*
 2002 Peer Review Committee, "Information Technology Research-Bioinformatics" (ITS2BI), National Science Foundation
 2002 Invited Speaker, Application of Bioinformatics in Cancer Detection Workshop, National Cancer Institution
 2002 Panel Member, BECON 2002: Sensor for Biological Research and Medicine, National Institutes of Health
 2002 *ad hoc* Study Section, "Development of Novel Technologies for in-vivo Imaging" (ZCA1-SRRB-9 (J2) (R)), National Institutes of Health

Honors

- 1999 Outstanding Faculty Research Achievement Award, The Catholic University of America

B. Selected Peer-Reviewed Publications

1. Y. Wang and J. M. Morris, "On Numerical Verification of Time-Domain Moment Method in Ultrasound Tomography," *SPIE Journal of Biomedical Optics*, vol. 1, no. 3, pp. 324-329, July 1996.
2. Y. Wang, T. Adali, and S-C B. Lo, "Automatic Threshold Selection Using Histogram Quantization," *SPIE Journal of Biomedical Optics*, vol. 2, no. 2, pp. 211-217, April 1997.
3. J. Zeng, Y. Wang, M. T. Freedman, and S. K. Mun, "Finger Tracking for Breast Palpation Quantification using Color Image Features," *SPIE Journal of Optical Engineering*, vol. 36, no. 12, pp. 3455-3461, December 1997.
4. Y. Wang, T. Adali, C-M Lau, and S-Y Kung, "Quantitative Analysis of MR Brain Image Sequences by Adaptive Self-Organizing Mixtures," *Journal of VLSI Signal Processing System*, vol. 8, pp. 219-239, 1998.
5. T. Adali, Y. Wang, N. Gupta, "A Block-wise Relaxation Labeling Scheme for Edge Detection in Cardiac MR Image Sequences," *International Journal of Imaging Science and Technology*, vol. 9, pp. 340-350, 1998.
6. Y. Wang, T. Adali, S-Y Kung, and Z. Szabo, "Quantification and Segmentation of Brain Tissues from MR Images: A Probabilistic Neural Network Approach," *IEEE Transactions on Image Processing*, vol. 7, no. 8, pp. 1165-1181, August 1998.
7. Y. Wang, S-H Lin, H. Li, and S-Y Kung, "Data Mapping by Probabilistic Modular Networks and Information Theoretic Criteria," *IEEE Transactions on Signal Processing*, vol. 46, no. 12, pp. 3378-3397, December 1998.
8. Y. Wang, L. Luo, M. T. Freedman, and S. Y. Kung, "Probabilistic Principal Component Subspaces: A Hierarchical Finite Mixture Model for Data Visualization," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 635-646, May 2000.
9. J. Xuan, T. Adali, Y. Wang, and E. Siegel, "Automatic Detection of Foreign Objects in Computed Radiography," *SPIE Journal of Biomedical Optics*, vol. 5, no. 4, pp. 425-431, 2000.
10. H. Li, Y. Wang, K-J R. Liu, S-H B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection-Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 289-301, April 2001.
11. H. Li, Y. Wang, K-J R. Liu, S-H B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection-Part II: Decision Support by Featured Database Visualization and Modular Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 302-313, April 2001.
12. Y. Wang, T. Adali, J. Xuan, and Z. Szabo, "Magnetic Resonance Image Analysis by Information Theoretic Criteria and Stochastic Site Models," *IEEE Transactions on Information Technology in Biomedicine*, vol. 5, no. 2, pp. 150-158, June 2001.
13. J. Xuan, T. Adali, Y. Wang, W. Hayes, J. Lynch, M. T. Freedman, and S. K. Mun, "A Computerized Simulation System for Prostate Needle Biopsy," *Simulation and Gaming*, vol. 32, no. 3, pp. 391-403, September 2001.
14. S-C B. Lo, H. Li, Y. Wang, and M. T. Freedman, "A Multiple Circular Path Neural Network Architecture for Detection of Mammographic Masses," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 150-158, February 2002.
15. Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, "Iterative Normalization of cDNA Microarray Data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 1, pp. 29-37, March 2002.
16. E. Matthew, Davis, N., Coop, A., Liu, M., Schumaker, L., Lee, R. Y., Srikanthana, R., Russell, C., Singh, B., Miller, W. R., Stearns, V., Pennanen, M., Tsangaris, T., Gallagher, A., Liu, A., Zwart, A., Hayes, D. F., Lippman, M. E., Y. Wang, and R. Clarke, "Development and Validation of a Method for Using Breast Core Needle Biopsies for Gene Expression Microarray Analysis," *Clinical Cancer Research*, vol. 8, pp. 1155-1166, May 2002.
17. Gu, Z., Lee, R. Y., Skaar, T., Bouker, K. B., Welch, J. N., Lu, J., Liu, A., Zhu, Y., Davis, N., Leonessa, F., Brünner, N., Y. Wang, and Clarke, R., "Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-6B, and cAMP response element binding with acquired resistance to Faslodex (ICI 182,780)," *Cancer Research*, vol. 62, pp. 3428-3437, June 2002.
18. Y. Wang, K. Woods, and M. McClain, "Information-Theoretic Matching of Two Point Sets," *IEEE Transactions on Image Processing*, vol. 11, no. 8, pp. 868-872, August 2002.
19. Z. Wang, J. Zhang, J. Lu, R. Lee, S-Y Kung, R. Clarke, and Y. Wang, "Discriminatory Mining of Gene Expression Microarray Data," *Journal of VLSI Signal Processing System*, 2002. in press

20. R. Srikanchana, J. Xuan, K. Huang, M. T. Freedman, C. Nguyen, and **Y. Wang**, "Non-rigid image registration by neural computation," *Journal of VLSI Signal Processing System*, 2002. in press

C. Research Support

Ongoing Research Support

R33 CA83231 Wang (PI) 7/1/99-8/31/03
 NIH/NCI
 Intelligent Mapping of Gene Expression Profiles
 Role: PI

DAMD17-98-8045 Wang (PI) 9/1/98-8/31/02
 DOD/CDMRP
 Improving Clinical Diagnosis by Change Detection in Image Sequences
 Role: PI

CUA-408217 Freedman (PI) 9/1/98-8/31/03
 Dues Technologies, Inc.
 Computer-aided Diagnosis for Lung Cancer Detection
 Role: Subcontract PI

DAMD17-01-0197 Liu (PI) 9/1/01-8/31/04
 DOD/CDMRP
 Computerized Tomography of Projection Ultrasound
 Role: Mentor

DAMD17-00-0195 Srikanchana (PI) 9/1/00-8/31/03
 DOD/CDMRP
 Tactile Imaging for Breast Cancer Diagnosis
 Role: Mentor

Completed Research Support

R21 RR12784 Wang (PI) 9/1/97-8/31/99
 NIH/NCRR
 Statistical Visualization of Localized Prostate Cancer
 Role: PI

R01 AG14400 Szabo (PI) 9/1/00-8/31/01
 NIH/NIAG
 Independent Component Imaging of Disease Signatures
 Role: Subcontract PI

CUA-408218 Wagner (PI) 9/1/00-8/31/01
 FDA/CDRH
 Multidimensional Receiver Operating Characteristics (ROC) Analysis for Computer-Aided Diagnosis
 Role: Subcontract PI

CUA-408211 Geng (PI) 9/1/97-8/31/01
 Genex Technologies, Inc.
 Volumetric Display Technologies
 Role: Subcontract PI

D. Book Chapters and Reviews

Y. Wang, T. Adali, and H. Li, "Neural Networks for Biomedical Signal Processing", *Handbook of Neural Networks for Signal Processing*, CRC Press, Y-H. Hu and J-N Huang editors, 2001.

Y. Wang and T. Adali, "Stochastic Model Based Image Analysis," *Signal Processing for Magnetic Resonance Imaging and Spectroscopy*, Marcel Dekker, H. Yan editor, 2002.

T. Adali and **Y. Wang**, "Image Analysis and Graphics for Multimedia Presentation," *Multimedia Image and Video Processing*, CRC Press, L. Guan, S-Y. Kung, and J. Larsen editors, 2000.

Y. Wang (invited), "Independent Component Analysis-A Book Review," *IEEE Transactions on Medical Imaging*, 2002. in press

E. Selected Proceedings Papers

1. **Y. Wang**, J. Zhang, K. Huang, J. Khan, and Z. Szabo, "Independent component imaging of disease signatures," *Proc. IEEE Intl. Symp. Biomed. Imaging*, July 7-10, Washington, DC 2002.
2. L. Lu, **Y. Wang**, Z. Wang, J. Xuan, S-Y. Kung, Z. Gu, and R. Clarke, "Discriminative mining of gene microarray data," *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 23-32, Falmouth, MA, September 2001.

Partially-Independent Component Analysis of Tumor Heterogeneities By DCE-MRI

Junying Zhang^a, Rujirutana Srikanthana^a, Jianhua Xuan^a, Peter Choyke^b,

King Li^b, and Yue Wang^a

^aDepartment of Electrical Engineering and Computer Science,
The Catholic University of America, Washington, DC 20064, USA
^bDiagnostic Radiology, National Institutes of Health, Bethesda, MD 20892, USA

ABSTRACT

Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) has emerged as an effective tool to access tumor vascular characteristics. DCE-MRI can be used to characterize microvasculature noninvasively for providing information about tumor microvessel structure and function (e.g., tumor blood volume, vascular permeability, and tumor perfusion). However, pixels of DCE-MRI represent a composite of more than one distinct functional biomarker (e.g., microvessels with fast or slow perfusion) whose spatial distributions are often heterogeneous. Complementary to various existing methods (e.g., compartment modeling, factor analysis), this paper proposes a blind source separation method that allows for a computed simultaneous imaging of multiple biomarkers from composite DCE-MRI sequences. The algorithm is based on a partially-independent component analysis, whose parameters are estimated using a subset of informative pixels defining the independent portion of the observations. We demonstrate the principle of the approach on simulated image data sets, and then apply the method to the tissue heterogeneity characterization of breast tumors. As a result, spatial distribution of tumor blood volume, vascular permeability, and tumor perfusion, as well as their time activity curves (TACs) are simultaneously estimated.

Keywords: Independent component analysis (ICA), partially-independent component analysis (PICCA), intrinsic dependency/non-intrinsic dependency of the components, dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), compartment model, time activity curves (TACs).

1. INTRODUCTION

Remarkable advances in functional imaging have been made in developing molecular-targeted contrast agents, ligands and imaging probes. Such imaging capabilities will allow for the visualization and elucidation of important disease-causing physiologic and molecular processes in living tissue. Subsequently, functional imaging will play an important role in the early detection, diagnosis, and treatment of diseases.¹ It is known that most advanced tumors are highly heterogeneous in structure that may reflect the underlying angiogenesis and/or metastasis.² Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) is a noninvasive imaging method for tumor microvascular characterization, which can be applied to assess (and potentially predict) the response to treatment including anti-angiogenic drugs. Kinetic characteristics changes following treatment have correlated with histopathological outcome (e.g., microvessel density) and patient survival. However, widespread success of DCE-MRI may be limited by the need for further technology development, particularly due to the lacking of quantitative and computational data analysis tools included by the instruments.

As a common problem in functional imaging, pixels represent a composite of more than one distinct molecular marker (i.e., the observed pixel intensity will consist of the weighted sum of activities of the various molecules). This problem exists for various reasons, e.g., target mixture, probe non-specificity, and kinetics or spectrum overlap. These aspects are briefly described as follows. First, mixed signals can result when distinct markers are

Further author information: (Send correspondence to Rujirutana Srikanthana)

E-mail: 55srikanchan@cua.edu, Telephone: 1 202 319 5243

Address: EE/CS Department, The Catholic University of America, Washington, D.C., U.S.A.

combined into a homogeneous mixture (e.g., fast and slow flow microvessels), independent of spatial resolution. Second, ligand-receptor binding depends largely on the three-dimensional shapes of both elements, where a ligand has many bonds that can be rotated into many different positions resulting in many shapes. Third, even with a precision excitation source, any overlap of the absorption spectra of the fluorophores leads to the excitation of multiple fluorophores whose emission spectra often also overlap. Thus, the observed signal intensity may well be composed of the emission from several markers of differing concentration and kinetics/spectrum (e.g., specific/nonspecific bindings, fast/slow flows). As a result, the overlap of multiple molecular signatures can severely decrease the sensitivity and specificity for the measurement of molecular signatures associated with different disease processes. As an example, imaging neuro-transporters in the brain requires the passage of radioligands across the blood brain barrier by ways of their high lipophilicity. But lipophilicity carries the risk of high nonspecific binding and retention in the white matter and could result in a bias of the estimated kinetic parameters that are used to measure binding to specific recognition sites.

It is well known that Independent Component Analysis (ICA)^{26,27} is a powerful method for blind source separation with a strong assumption that the sources are independent to each other. This paper describes a computation approach to dependent component imaging, where functional imaging is the case. The method is to identify an informative index subspace and over which to separate mixed imagery sources by partially-independent component analysis (PICA), whose parameters are estimated using informax principle. We discuss the theoretic roadmap of the approach, and its applications to computer simulation phantoms and DCE-MRI sequences of breast tumor.

2. THEORY AND METHOD

Independent component analysis (ICA)²⁶ is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. The application of ICA has been found in many separate fields such as feature extraction, image processing, medical image processing, telecommunication, econometric signal processing, and so fourth.^{11,27} The method aims at recovering the unobservable independent sources (or signals) from multiple observed data masked by linear or nonlinear mixing of the components.²⁷ One of the basic assumptions for ICA model is the statistical independence between components.²⁶ However, the dependent components are often occurred in the real world situation, including functional imaging derived from tissue samples.

2.1. Compartment Modeling

Compartment modeling forms the basis for tracer characterization in DCE-MRI.³ Fig. 1 shows a parallel mode two-tissue compartment model.² The conventional compartment model leads to a set of first order differential equations:

$$\begin{aligned}\dot{c}_f(t) &= k_{1f}c_p(t) - k_{2f}c_f(t) \\ \dot{c}_s(t) &= k_{1s}c_p(t) - k_{2s}c_s(t) \\ c_t(t) &= c_f(t) + c_s(t) \\ c_m(t) &= c_f(t) + c_s(t) + c_p(t)\end{aligned}\tag{1}$$

where $c_f(t)$ and $c_s(t)$ are the tissue activity in the fast turnover and slow turnover pools, respectively, at time t ; $c_p(t)$ is the tracer concentration in plasma (i.e., the input function); $c_t(t)$ is the total tissue activity; $c_m(t)$ is the measured total tissue activity; k_{1f} and k_{1s} are the unidirectional transport constants from plasma to tissue (ml/min/g: spatially shift-varying); and k_{2f} and k_{2s} are the rate constants for efflux (/min: spatially shift-invariant). It is important to note that $c_f(t)$, $c_s(t)$, $c_p(t)$, and $c_m(t)$ are also called the *time-activity curves* (TACs) associated with a pre-defined region of interest (ROI).

It can be shown that $c_f(t)$ and $c_s(t)$ can be solved analytically in a parametric form

$$\begin{aligned}c_f(t) &= k_{1f}c_p(t) \otimes e^{-k_{2f}t} \\ c_s(t) &= k_{1s}c_p(t) \otimes e^{-k_{2s}t}\end{aligned}\tag{2}$$

where \otimes denotes the mathematical convolution operation. By fitting $c_m(t)$ to the measured ROI TAC in the light of pre-acquired $c_p(t)$, the model parameters (k_{1f} , k_{1s} , k_{2f} , k_{2s}) can be estimated.^{3,4}

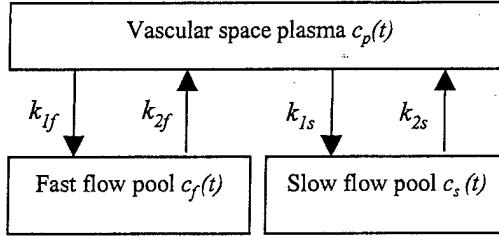


Figure 1. Two-tissue compartment model (parallel mode).

Based on linear system theory, a simple method can be developed to convert temporal kinetics to spatial information.^{2,5} First, we can normalize the ROI based tissue kinetics to define three TACs for each pixel

$$\begin{aligned} a_f(t) &= \frac{c_f(t)}{k_{1f}} = c_p(t) \otimes e^{-k_{2f}t} \\ a_s(t) &= \frac{c_s(t)}{k_{1s}} = c_p(t) \otimes e^{-k_{2s}t} \\ a_p(t) &= \frac{c_p(t)}{v_p} \end{aligned} \quad (3)$$

where v_p is the plasma volume in tissue. Second, for pixels $i = 1, \dots, N$ within an ROI, we let $k_{1f}(i)$ and $k_{1s}(i)$ be the local model parameters and use them to describe the dynamics of each pixel in the ROI

$$c_m(i, t) = k_{1f}(i)a_f(t) + k_{1s}(i)a_s(t) + v_p(i)a_p(t) \quad (4)$$

where $c_m(i, t)$ is the measured pixel TAC, $k_{1f}(i)$ and $k_{1s}(i)$ are the permeability of fast and slow turnover regions in the pixel, respectively, and $v_p(i)$ is the plasma volume in the pixel. We call this representation as factored compartment modeling.⁵

Third, let (t_1, t_2, \dots, t_n) be the sampling time points of the DCE-MRI measurements. Then, the linear least square solution of Eq. (4) can be given by the following equation:

$$\begin{bmatrix} k_{1f}(i) \\ k_{1s}(i) \\ v_p(i) \end{bmatrix} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \begin{bmatrix} c_m(i, t_1) \\ c_m(i, t_2) \\ \vdots \\ c_m(i, t_n) \end{bmatrix} \quad (5)$$

where

$$\mathbf{A} = \begin{bmatrix} a_f(t_1) & a_s(t_1) & a_p(t_1) \\ a_f(t_2) & a_s(t_2) & a_p(t_2) \\ \vdots & \vdots & \vdots \\ a_f(t_n) & a_s(t_n) & a_p(t_n) \end{bmatrix} \quad (6)$$

The estimated values of $k_{1f}(i)$, $k_{1s}(i)$ and $v_p(i)$ vary from pixel to pixel and reconstruct three factor images respectively. In particular, factor images $k_{1f}(i)$ and $k_{1s}(i)$ represent ROI sub-regions with fast and slow kinetics, respectively.

Preliminary effort has been recently made to perform blind compartment modeling without any knowledge of the input function.¹²⁻¹⁵ An initial effort, eigenvector based multichannel blind deconvolution (EVAM),¹⁶ was used to estimate the parameters of a two-tissue compartment model for PET FDG imaging,¹² but was shown to give relatively poor (sensitive to noise) and non-unique estimates in a simulation study.¹³ A more optimal solution was proposed on the application of the iterative quadratic maximum-likelihood (IQML) method to parameter estimation.¹³ The blind identification problem is treated as a nonlinear least square problem whose variables are separate.¹⁷ Other approaches in which both the input function and kinetic parameters are treated as unknowns have been explored in [14, 15].

3. PARTIALLY-INDEPENDENT COMPONENT ANALYSIS (PICA)

3.1. Independent Component Analysis (ICA)

As aforementioned, one potential limitation associated with compartment analysis is that they are all restricted to a parametric (thus simplified) model that may not adequately describe the underlying physiological or biochemical processes about tracer-target interactions, in addition to the likely invasive acquisition of the input function. Although factor analysis (FA) attempts to solve the problem, the results were mostly unsatisfactory.

From linear system theory,²⁴ it can be shown that the solution (zero-state response) to a kinetic system has the very general form as shown in Eq. (4), or Eq. (7) as represented in vector-matrix form. This motivates the consideration of a statistically-principled computational approach involving newly invented independent component analysis (ICA) theory.^{11, 26, 27} The goal is to blindly and computationally reconstruct both \mathbf{A} and \mathbf{k} based on \mathbf{c}_m . This philosophy for computed simultaneous imaging of multiple biomarkers is similar in spirit to the blind source separation (BSS) for solving the *cocktail-party problem*.²⁶

From latent variable model interpretation,²⁸ Eq. (7)

$$\begin{bmatrix} c_m(i, t_1) \\ c_m(i, t_2) \\ \vdots \\ \vdots \\ c_m(i, t_n) \end{bmatrix} = \mathbf{A} \begin{bmatrix} k_{1f}(i) \\ k_{1s}(i) \\ v_p(i) \end{bmatrix}, \quad (7)$$

describes how the observed data are generated by a process of mixing the latent (or “hidden”) variables, where matrix \mathbf{A} is called the *mixing matrix*, the factor images (or “source signals”) are not observable, and nothing is known about the properties of the TACs (or “mixing process”). In the absence of this information, one has to proceed “blindly” to recover the factor images from their TAC-modulated activity mixtures.⁵

We can state such *computed simultaneous imaging of multiple biomarkers* as follows: “Given N independent realizations of the measured pixel TAC vector $\mathbf{c}_m(i, t)$, $i = 1, 2, \dots, N$, find an estimate of the inverse of the TAC-mixing matrix $\mathbf{A}(t)$ and factor image vector $\mathbf{k}(i) = [k_{1f}(i), k_{1s}(i)]^T$.”

ICA method, as a newly invented statistical and neural computation technique, promises a powerful computational tool for separating hidden sources from mixed signals when many classic methods fail completely.²⁶ ICA method utilizes *independence* as a guiding principle and performs BSS based on a *nongaussian* factor analysis with a unique solution.¹¹ More precisely, by assuming that the hidden components are statistically independent with nongaussian distributions, these hidden sources can be found by ICA, except for an arbitrary scaling of each signal component and permutation of indices. In other words, it is feasible to find a demixing matrix \mathbf{W} whose individual rows are a rescaling and permutation of those of the mixing matrix \mathbf{A} . ICA approach exploits primarily *temporal diversity* in that the dynamic images taken at different times carry different mixtures of the factor images.⁵ There are several algorithms for ICA that are derived from different optimization principles. More details can be found in [11, 26].

3.2. Partially-ICA

We have found that direct application of ICA to tumor heterogeneity characterization using all the pixels, however, often leads to an unsatisfactory recovery of factor images $\mathbf{k}(i)$. By a closer look at the joint distribution of the factor images, we found that they are often not statistically independent over the whole pixel set.⁵ This shall not be a surprise since factor images are expected to be piece-wise continuous thus form clusters over the joint distribution. It can be further concluded that such joint distribution clusters correspond to the overlapped homogeneous areas of the factor images. Thus, we shall expect to achieve a better factor image decomposition using a subset of pixels that supports the independency of the factor images.

Inspired by such reasoning, we proposed a partially-ICA (PICA) technique in [5]. Rather than using all the pixels that give rise to a large decomposition error due to source dependency, we attempt to (iteratively)

identify a pixel subset supporting source independency and over which to estimate the demixing matrix \mathbf{W} and subsequently factor images $\mathbf{k}(i)$.

Compared with the basic ICA model, where each observation is a linear combination of independent components, our PICa model assumes that each observation x_i is a linear mixture of statistically dependent components s_1, s_2, \dots, s_n , with an n by n non-singular mixing matrix A , i.e.,

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \mathbf{A} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix}, \text{ or } \mathbf{x} = \mathbf{As} \quad (8)$$

where

$$\mathbf{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1m} \\ s_{21} & s_{22} & \dots & s_{2m} \\ \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & \dots & s_{nm} \end{bmatrix} = [S_1, S_2, \dots, S_m], \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} = [X_1, X_2, \dots, X_m] \quad (9)$$

and m is the number of pixels in each functional image. Our task is to recover the dependent components from observations by still utilizing ICA.

Let us review the mechanism of independent component analysis on basic ICA model. The components S_i are statistically independent, while based on the Central Limit Theorem; the distribution of a sum (observations) of independent random variables (components) tends toward a Gaussian distribution, under certain conditions. Therefore, the procedure that ICA searches for estimates of the components is to find directions, such that the projections of observations on each direction are distributed with most non-Gaussian distribution. This is the reason that ICA algorithm can help find the (statistically) precise estimate of the components if the components are completely independent (except those two ambiguities of ICA on the scale and order of the components). However, it will mislead direction finding if the components are dependent but ICA algorithm is still superimposed on the observations, which are mixtures of the dependent components. Also, it will give rise to a large separation error since all the pixels are utilized for ICA calculation while the components over all these pixels are statistically dependent.

In order to effectively utilize ICA for this problem, an informative pixel index subspace (corresponding to the independent part of the components) needs to be identified. Then we can perform ICA over this subspace to recover the estimated mixing matrix over the subspace. Imposing this mixing matrix over all the pixel indices, the dependent components are then available to be recovered. The key point is to identify the informative pixel indices, over which the components are statistically independent. The difficulty of this approach is that the independent subspace of the components needs to be identified without any statistical information derived from the components themselves: the components are just what need to be recovered. We only have information derived from observations rather than from components. For simplicity, we consider the dependency of the components rather than that of the observations. It is well known that the statistical dependency/independency of the components s could be measured (visualized) by means of scatter plot with m sample points S_1, S_2, \dots, S_m in n dimensional space (each dimension corresponds to a component): they consist of dependent/independent components, which we can perform linear/nonlinear regression curve to estimate the statistical relation between n components.

Clearly, when the number of pixels m is much larger than the number of components n , which is the situation for functional imaging, most probably the sample points are clustered into some clusters. Based on this observation, we divide the dependency of the components into two categories: intrinsic dependency and non-intrinsic dependency. By intrinsic dependency of the components, we mean the dependency caused by the linear and/or nonlinear correlation between components over cluster centers. We refer to non-intrinsic dependency of the components as the dependency over sample points inside each cluster. In other words, the intrinsic dependency corresponds to the global dependency among clusters, while non-intrinsic dependency corresponds to the local

dependency among samples in each cluster. For ICA to work well, we need to remove both intrinsic dependency and non-intrinsic dependency of the components. This can be accomplished by removing the pixels that contribute to intrinsic dependency and non-intrinsic dependency of the components, while retaining the informative pixel indices. Finally, we will apply the ICA onto the subset of pixels for effective recovery of the components over those informative pixel indices.

Notice that both intrinsic dependency and non-intrinsic dependency of the components should be removed with only the information from observations rather than from components. A possible approach to remove intrinsic and non-intrinsic dependencies can be summarized as follows. The intrinsic dependency of the components can be removed by removing all of the clusters except one. The non-intrinsic dependency of the components can be alleviated by removing some sample points in the remained cluster. We will describe the removal procedures next.

3.2.1. Intrinsic dependency removal

In order to remove intrinsic dependency of the components from the observations, the joint distribution of the observations is estimated with Expectation-Maximization (EM) algorithm initialized by k-means method. We assume that the number of clusters t is known with some prior knowledge of the problem. In our DCE-MRI study of breast tumor, the prior knowledge tells us that this number should be two considering the fast-flow and slow-flow characteristics of the breast tumor. Assume that the distribution of the observations is in the following form:

$$p(x) = \sum_{i=1}^t \pi_i p(x; m_i, \sigma_i^2) \quad (10)$$

where $p(x; m_i, \sigma_i^2)$ is the i th Gaussian distribution with m_i and σ_i^2 as its mean and variance, and π_i is the weight of the i th Gaussian distribution, $\sum_{i=1}^t \pi_i = 1$. Notice that each Gaussian distribution forms a cluster in scatter plot of the observations.

We remove intrinsic dependency of the components in a statistical way. Each sample point X_j is removed in observation scatter plot statistically, with the probability of

$$p_{r1}(\text{removal of } X_j) = \frac{p(X_j) - \pi_k p(X_j; m_k, \sigma_k^2)}{p(X_j)} \quad (11)$$

where k is the index of the remaining cluster.

3.2.2. Non-intrinsic dependency Removal

For removal of non-intrinsic dependency of the components, the remained sample points in scatter plot of the observations are statistically down-sampled with P_θ as its parameter, i.e., X_j is removed statistically with the probability of

$$p_{r2}(\text{removal of } X_j) = \begin{cases} \pi_k p(X_j; m_k, \sigma_k^2) - P_\theta & , \text{if } \pi_k p(X_j; m_k, \sigma_k^2) \geq P_\theta \\ 0 & , \text{otherwise} \end{cases} \quad (12)$$

As a result, the remained sample points in scatter plot of the observations would be uniformly distributed. It is important that the statistical dependency of the components over the corresponding sample points are removed without any destruction of the linear dependency of the observations over the remained sample points. In fact, this dependency should not be removed for the following ICA computation, because the observations over the remained sample points are still the linear mixing of the components, which are statistically independent over the corresponding sample points from the PICA model.

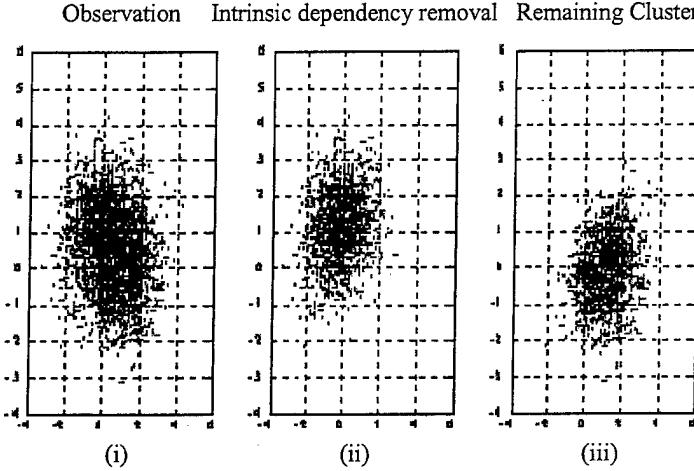


Figure 2. Removal of the intrinsic dependency between the components: (i) observation consists of more than one cluster; (ii) the removal of intrinsic cluster dependency; (iii) the remaining cluster after the removal of intrinsic dependency.

3.2.3. Proposed Algorithm

Assume I is the remained sample point index subset, which is obtained by the removal of intrinsic dependency and non-intrinsic dependency of the components from the scatter plot of the observations. I corresponds to the independent pixel index subspace of the components. We have following procedure for a composite separation of observations:

Step 1: estimate the joint distribution $p(x)$ of the observations with EM algorithm initialized by k-means method, where the number of clusters t is assumed to be known from some prior knowledge;

Step 2: remove intrinsic dependency of the components by a statistical cluster removal method with eq. 11, and remove non-intrinsic dependency of the components by a statistical down-sampling method by utilizing eq. 12 with parameter P_θ , both in the scatter plot of the observations; assume the retained pixel index subset is I ;

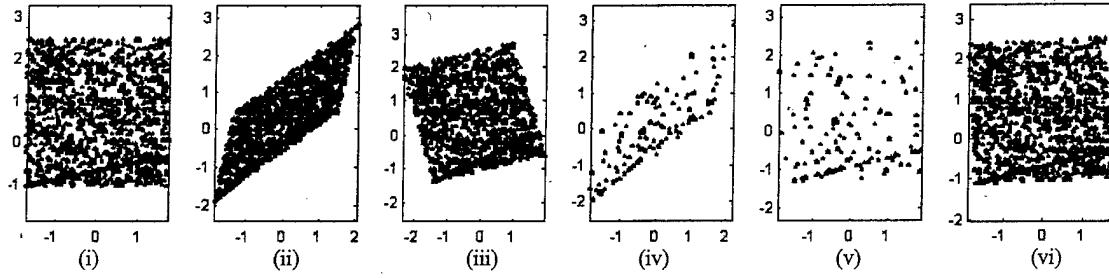
Step 3: perform ICA on the retained part of the observations, $\{X_j, j \in I\}$, to obtain the estimated mixing matrix A . Since the observations over the retained pixel index subspace I are independent, which satisfies the assumption of basic ICA model; the mixing matrix A is expected to be better estimated except the ambiguity of scaling and ordering.

Step 4: impose the estimated mixing matrix on the observations over entire indices to obtain the estimated components, i.e., $s = A^{-1}x$.

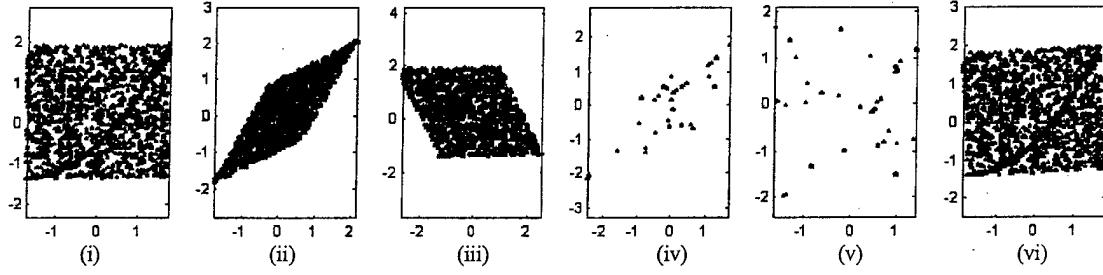
There are two parameters, t and P_θ , in the above algorithm. t is set with some number by prior knowledge of the problem, while P_θ is a trade-off parameter for controlling the uniformity of the distribution over the sparseness of the sample points in the scatter plot of observations after the removal procedures. The smaller the P_θ is, the more uniform the distribution of sample points over the index subspace looks like, and the less sample points over the index subspace will be. The directions searched out for the sample points over the index subspace with ICA algorithm determine the rows of the estimated mixing matrix. Theoretically speaking, if the parameter P_θ is set smaller, the more uniformed distribution of the sample points in scatter plot of the observations over the index subspace will merit this direction searching; On the other hand, the number of sample points becomes less that limits the sample points to form that distribution, demeriting this direction searching.

4. RESULT AND DISCUSSION

We first applied our method to two computer simulation phantoms for the removal of intrinsic dependency and non-intrinsic dependency among components. Then we applied our method to a data set generated by compartment models, and a real DCE-MRI data set of tumor heterogeneity study.



(a) The situation of linear dependency between components



(b) The situation of non-linear dependency between components

Figure 3. The scatter plots of the component recovery process, for both linear (a) and non-linear (b) dependency cases, with ICA and PICA method, (i) components, (ii) observations, (iii) recovered components from ICA method, (iv) observations over the informative index subspace, (v) recovered components over the informative index subspace, (vi) recovered components over full index space from PICA method.

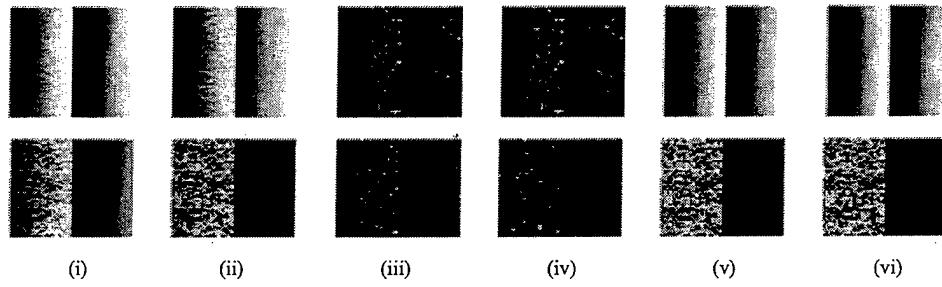
4.1. Phantom study

Figure 2(i) shows the data set that is generated by a sum of two Gaussian distributions: centered at $(0,1)$ and $(1,0)$ with standard variations as two components. Clearly, the components are not independent (i.e., they are intrinsically dependent), and the removal of either Gaussian distribution makes the sample points over the index subspace statistically independent (see Figure 2(ii) and (iii)).

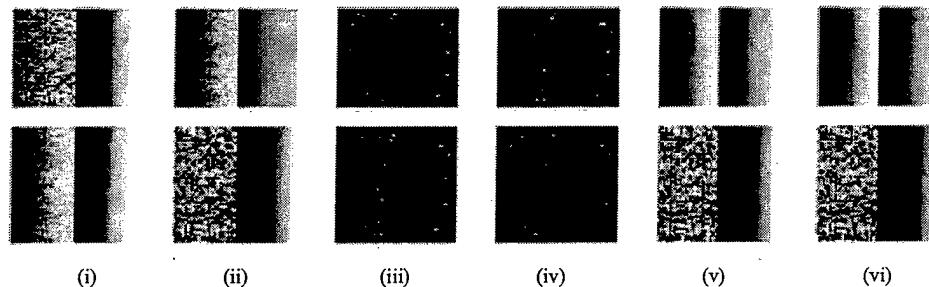
The next data set is generated such that the two components are independent in their first half (left half), and have linear/non-linear correlation for the other half (right half), both with uniformly distributed intensities. The mixing matrix is randomly generated to form observations. Clearly there is a linear/non-linear non-intrinsic dependency between the components in this experiment. Figure 3 (upper/lower figure) shows the scatter plots in the component recovery process with ICA and PICA methods for the situation of linear/non-linear dependency between the components. Figure 4 shows the component recovery process, where the randomized intensities of the first half and the second half are shown in two-dimensional images. Note that the pixels for the first component in the figure are reordered according to its intensities. Evidently, the estimated components recovered by PICA method are much closer to the ground truth of the components by comparing both the scatter plots in Figure 3(i), (iii), (vi) and the images in Figure 4(ii), (iv), (vi).

4.2. Experiments on compartment model

A compartment model is used to simulate the dynamic behavior of the breast tumor obtained with a DCE-MRI sequence. The mask for the fast-flow (FF)/slow-flow (SF) patterns, as well as the overlap region of FF and SF, is shown in Figure 5(a). The pixel intensity in the FF-dominant region/overlap region/SF-dominant region is a linear combination of FF and SF patterns with the corresponding intensity weights of $0.9/0.5/0.1$ and $0.1/0.5/0.9$. In Figure 5(a), bright/dark gray color corresponds to FF/SF region, and the light gray/dark corresponds to overlap/background, respectively. From the compartment model, we can get a sequence of images shown in Figure 5(b). Our task is to identify the FF and SF patterns hidden in the observed sequence of images.



(a) The situation of linear dependency between components



(b) The situation of non-linear dependency between components

Figure 4. Recovery results from ICA and PICA, (i) observations, (ii) recovered components by utilizing ICA method, (iii) observations over the informative index subspace, (iv) recovered components over the informative index subspace, (v) recovered components over full index space by utilizing PICA method, (vi) real components.

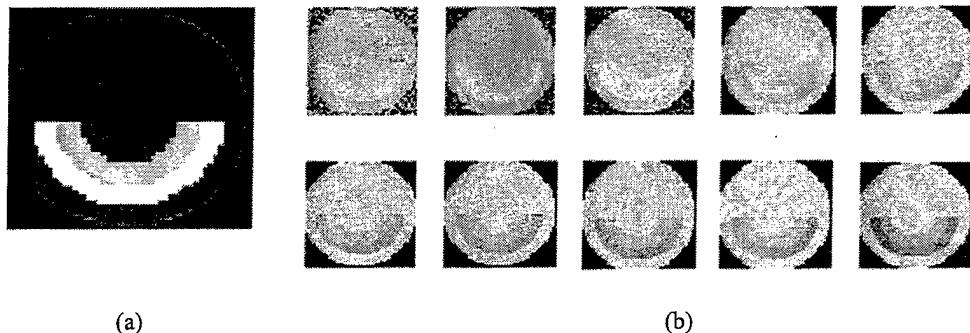


Figure 5. A simulator tumor phantom including fast and slow kinetic subregions; (a) Image mask in compartment model, (b) The sequence of images from the compartment model with the mask shown in (a)

We applied our PICA method to separate the observed composite images into two FF and SF patterns, as well as to have the time-activity curve (TAC). The only two spatial patterns, FF and SF, which is a prior knowledge of this problem, induced us to make an assumption of two components from the observed image sequence. Thus, we divided the image sequence into the first half of the sequence and the second half of the sequence, according to the time index. Then PICA was performed for each pair of observed images, one from the first half, and another from the second half, and the corresponding components were estimated. The final estimation of the components (spatial FF/SF patterns) was attained by averaging all of the estimated components obtained from each pair of observed images. Figure 6 shows the spatial FF/SF pattern and the corresponding TACs achieved by PICA, together with those by ICA for comparison. Notice that each TAC according to the compartment model should be a positive curve. Figure 6(a) shows a better performance for the estimation of the FF and

SF patterns, which are closer to the ground truth of 0.9:0.5:0.1 and 0.1:0.5:0.9/0.9:0.3:0.1 and 0.1:0.7:0.9 from Figure 6(a)(i)/(ii). It is shown from Figure 6(b) that TACs from PICA are better than those from ICA, since they are both approximately positive and more fit to the meaning of dynamic fast-flow and slow-flow activity of the tumor tissue.

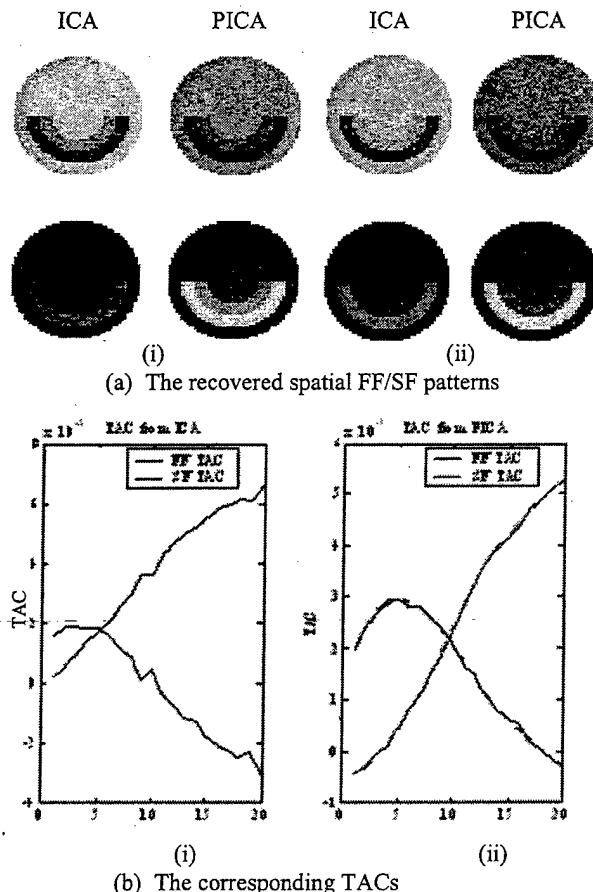


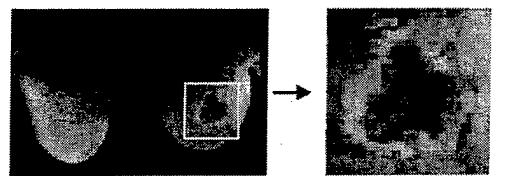
Figure 6. ICA and PICA result, (a) the factor images results for compartment model with FF:overlap:SF region intensity for FF pattern and for SF pattern to be 0.9:0.5:0.1 and 0.1:0.5:0.9/0.9:0.3:0.1 and 0.1:0.7:0.9 respectively; (b) the corresponding TACs.

4.3. Experiment on real DCE-MRI data set

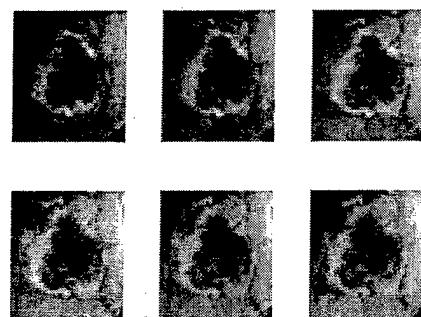
We tested our PICA method with a real DCE-MRI sequence of breast tumor studies. Figure 7 shows a typical sequence of breast tumor DCE-MRI study. Compared with results from the direct application of ICA, our results using PICA shown in Figure 8, were quite promising in the extracted factors that closely resemble the expected characteristics of compartmental kinetics of tumors. The factor images and the TACs reveal regional distribution of the FF and SF patterns.

ACKNOWLEDGMENTS

This work was supported by the US Army Medical Research and Materiel Command under Grants DAMD17-00-0195 and DAMD17-98-8045.



(a) Tumor (breast cancer) spatial heterogeneity revealed by DCE-MRI.



(b) Typical sequences of DCE-MRI of breast tumor

Figure 7. (a) Tumor (breast cancer) spatial heterogeneity revealed by DCE-MRI; (b) Sequences of the breast tumor ROI.

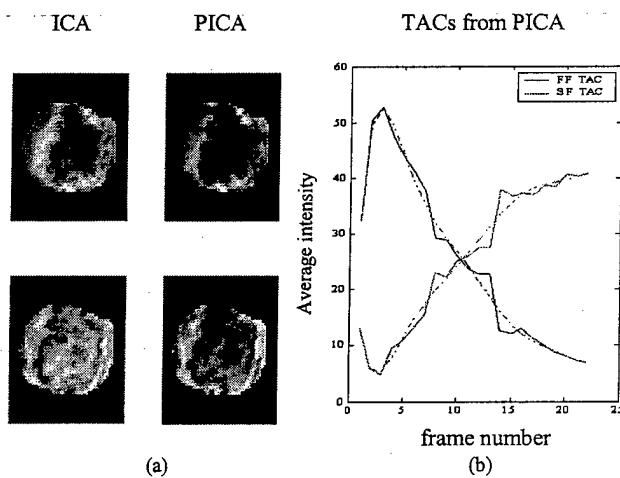


Figure 8. The recovered factor images of the breast cancer spatial heterogeneity and the corresponding TACs revealed by DCE-MRI with PICA method

REFERENCES

1. Z. Szabo, P. F. Kao, W. B. Mathews, H. T. Ravert, J. L. Musachio, "Positron Emission Tomography of 5-HT Reuptake sites in the human brain with C-11 McN5652 Extraction of characteristic images by artificial neural network analysis," *Behavioral Brain Research*, vol.73, pp.221-224, 1996.
2. Y. Zhou, S-C Huang, T. Cloughesy, C. K. Hoh, K. Black, and M. E. Phelps, "A modeling-based factor extraction method for determining spatial heterogeneity of Ga-68 EDTA kinetics in brain tumors," *IEEE Trans. Nuclear Sci.*, vol. 44, no. 6, pp. 2522-2527, Dec. 1997.

3. R. N. Gunn, Steve R. Gunn, and V. J. Cunningham, "Positron emission tomography compartmental models," *J. Cerebral Blood Flow and Metabolism*, vol. 21, no. 6, pp. 635-652, 2001.
4. PKIN: Kinetic Modeling, <http://www.pmod.com/doc/PkinReference.html>.
5. Y. Wang, J. Zhang, K. Huang, J. Khan, and Z. Szabo, "Independent component imaging of disease signatures," *Proc. IEEE Intl. Symp. Biomed. Imaging*, pp. 178-181, July 7-10, Washington, DC 2002.
6. W. J. Geckle and Z. Szabo, "Physiologic factor analysis (PFA) and parametric imaging of dynamic PET images," *IEEE Symposium on Medical Imaging*, 1992.
7. L. Cinotti, J. P. Bazin, R. DiPaola, H. Susskind, and A. B. Brill, "Processing of Xe-127 regional pulmonary ventilation by factor analysis and compartmental modeling," *IEEE Trans. Med. Imaging*, vol. 10, no. 3, pp. 437-444, 1991.
8. J. Y. Ahn, D. S. Lee, J. S. Lee, S. K. Kim, G. J. Chon, J. S. Yeo, S. A. Shin, J. K. Chung, and M. C. Lee, "Quantification of regional myocardial blood flow using dynamic H215O PET and factor analysis," *J. Nuclear Med.*, vol. 42, no. 5, pp. 782-787, May 2001.
9. H. H. Harman, *Modern Factor Analysis*, University of Chicago Press, 2nd Ed., 1967.
10. M. Samal, H. Surova, M. Karny, et al., "Enhancement of physiological factors in factor analysis of dynamic studies," *Eur J. Nucl. Med.*, 12: 280-283, 1986.
11. H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803-851, 1999.
12. C-H Lau, P-K D. Lun, and D. Feng, "Non-invasive quantification of physiological processes with dynamic PET using blind deconvolution," *Proc. ICASSP*, vol. 3, pp. 1805-1808, Seattle, WA 1998.
13. D. Y. Riabkov and E. V. R. Di Bella, "Estimation of kinetic parameters without input functions: analysis of three methods for multichannel blind identification," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 11, pp. 1318-1327, Nov. 2002.
14. D. Feng, K-P Wong, C-M Wu, and W-C Siu, "A technique for extracting physiological parameters and the required input function simultaneously from PET image measurements: theory and simulation study," *IEEE Trans. Info. Tech. Biomed.*, vol. 1, no. 4, pp. 243-254, Dec. 1997.
15. K-P Wong, D. Feng, S. R. Meikle, and M. J. Fulham, "Simultaneous estimation of physiological parameters and the input function-in vivo PET data," *IEEE Trans. Info. Tech. Biomed.*, vol. 5, no. 1, pp. 67-76, Mar. 2001.
16. M. I. Gurelli and C. L. Nikias, "EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *IEEE Trans. Signal Processing*, vol. 43, pp. 134-149, Jan, 1995.
17. L. Tong and S. Perreau, "Multichannel blind identification: from sub-space to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951-1968, Oct. 1998.
18. P. Santiago and H. D. Gage, "Quantification of MR brain images by mixture density and partial volume modeling," *IEEE Trans. Med. Imaging*, vol. 12, no. 3, pp. 566-573, 1993.
19. P. Santiago and H. D. Gage, "Statistical models of partial volume effect," *IEEE Trans. Image Processing*, vol. 4, no. 11, pp. 1531-1540, 1995.
20. P. Tamayo, D. Slonim, J. Mssirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci.*, vol. 96, pp. 2907-2912, March 1999.
21. D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical analysis of finite mixture distributions*. New York: John Wiley, 1985.
22. Y. Wang, L. Luo, M. T. Freedman, and S-Y Kung, "Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization," *IEEE Trans. Neural Nets*, Vol. 11, No. 3, pp. 625-636, May 2000.
23. Z. Wang, J. Zhang, J. Lu, R. Lee, S-Y Kung, R. Clarke, and Y. Wang, "Discriminatory mining of gene expression microarray data," *Journal of VLSI Signal Processing System*, 2003. in press
24. W. J. Rugh, *Linear System Theory*, Prentice-Hall, Inc., 1996.
25. Y. Wang, "Independent component analysis-a book review," *IEEE Trans. Med. Imaging*, vol. 21, no. 7, pp. 839-840, July 2002.
26. A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: John Wiley, 2001.
27. S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Ney Jersey: Prentice-Hall, 1999.
28. B. S. Everitt, *An Introduction to Latent Variable Models*, London: Chapman and Hall, 1984.